ESTADISTICA APLICADA

CON

SPSS

(Descriptiva Inferencia)

MARIO BLACUTT MENDOZA

Los derechos de autor de las versiones impresa y digital de la presente obra están debidamente reservados y protegidos por Ley

La Paz, Bolivia 2000

PRÓLOGO

Este manual está dirigido, principalmente, a los profesionales, estudiantes y hombres de negocios que necesiten un instrumento aplicable en todas las áreas del conocimiento. En particular, a economistas, administradores de empresas, psicólogos, sociólogos y, en general, a todas las personas que deseen contar con dos disciplinas expresadas en una versión ágil y oportuna para recopilar, organizar, manipular, explicar, pronosticar e interpretar datos y convertirlos en información adecuada al proceso de tomar decisiones.

En los primeros capítulos analizaremos algunos conceptos fundamentales en la estructura de la Estadística, para ir consolidando el uso de los mismos a medida que ingresemos en etapas más avanzadas, todo ello, de una manera que resultará muy asequible a cada uno de los participantes. Luego nos trasladaremos a la Inferencia Estadística. El método es muy sencillo: en cada capítulo se explicará, con los detalles necesarios, el significado de los conceptos estadísticos correspondientes, la manera de usarlos, su utilidad y su interpretación. En el proceso, tendremos la gran ayuda del programa estadístico SPSS, siglas en inglés de su nombre completo: *Scientific Program for Social Sciences,* que es el más conocido y usado de todos los paquetes estadísticos modernos. Así, haremos algo que generalmente requiere dos cursos diferentes: aprenderemos Estadística Descriptiva y el SPSS.

Así, el Manual está dividido en dos partes principales

Todos los pasos mostrarán la conexión unitaria entre el aprendizaje de la Estadística con el uso de los programas del SPSS para resolver cada tipo de problemas. En la actualidad, las principales empresas terciarizan los trabajos de investigación de datos; sin embargo, para cumplir esa tarea, sólo necesitamos conocer la esencia de los conceptos estadísticos, operar el SPSS e interpretar sus resultados.

La Metodología

Todo lo que digamos sobre Estadística será concretado de inmediato con la manera de usar el SPSS y mostrar cómo se aplica a la realidad mediante el uso de la computadora y el SPSS. En virtud de que se trata de un Manual de Estadística Aplicada orientado a resolver los problemas prácticos que se presentan todos los días, no habrá demostraciones matemáticas. Los cálculos y toda la operatividad serán realizados por el SPSS, cuyos programas sí, han sido estructurados sobre la base de las fórmulas matemáticas requeridas. Por lo general, el aprendizaje de la estadística en los cursos universitarios se hace innecesariamente artificial y difícil, pues se asume que todos los participantes estudian para obtener el título de Estadísticos Teóricos. En los textos tradicionales, las demostraciones teóricas de las fórmulas ocupan la mayor parte de los capítulos y los participantes pierden el rumbo, confundiendo el concepto cualitativo con el proceso de la demostración. Este manual está diseñado para brindar al participante un instru-

mento ágil y operativo que le sirva para resolver problemas reales en entornos reales. De ahí el nombre: Estadística Aplicada con el SPSS.

Por último, me gustaría referirme al modo expositivo que caracteriza el desarrollo del curso; he tratado de que sea conciso y claro.

He puesto especial énfasis en que nada esté por demás ni por de menos.

Breve esbozo sobre la Historia de la Estadística

Los eruditos, esos señores que tienen la mitad del conocimiento en sus bibliotecas y la otra mitad en sus cerebros, dicen que la Estadística surgió como un instrumento de análisis en Egipto, por el celo de las autoridades en conocer la población, la cantidad de tierra disponible, los repartos de esa tierra y la riqueza que poseían; pero, sobre todo, para obtener la información necesaria al cálculo de los impuestos. Los chinos ya conocían sobre la técnica de levantar censos y los griegos no se quedaban atrás. Los romanos asimilaron el conocimiento anterior al que le sumaron sus propios descubrimientos en la tarea periódica de levantar censos, en los que se incluía datos sobre las cabezas de ganado, los recursos naturales, como también, los matrimonios, nacimientos y defunciones. La Edad Media no trajo nada nuevo, pero el Renacimiento sí, fue una época en la que se dio gran importancia a las técnicas de recopilar, ordenas e interpretar datos, que es la médula de la Estadística.

En la primera mitad del siglo XVI, los alemanes hicieron una recopilación sobre los recursos naturales, la población y otros similares. Por aquellas épocas había una creencia muy difundida en sentido de que en los años terminados en 7, el número de muertos era mucho mayor que en los demás. Gaspar Neumann, un científico de gran voluntad y conocimiento se dio a la tarea de revisar las partidas parroquiales para comparar el número de nacimientos y defunciones de decenas de años. Su investigación le permitió negar la fatídica sombra de los años terminados en 7.

Como sucede con todo descubrimiento útil a la ciencia, los métodos usados por Neumann se expandieron. Un astrónomo inglés los leyó con gran atención, los interpretó debidamente y los enriqueció con sus propias ideas. Todo eso permitió que el actual cometa Halley llevara su nombre. Además, usó de los métodos estadísticos para sentar las bases que sustentan la estructura de lo que ahora se denominan "Tablas de Mortalidad", médula espinal de las compañías de seguros.

En Grecia, la primera referencia con relación a la futura disciplina estadística podría ser la que se incluye en el Libro II de Tucidides sobre la Guerra del Peloponeso entre espartanos y atenienses. En el texto se anota conceptos propios de lo que ahora llamamos *Muestreo*. El problema, al parecer, fue el siguiente. El ejército debe asaltar una muralla y los jefes militares han decidido que es preciso contar con una torre móvil que permita a los soldados tomar la ciudadela, minimizando el riesgo. La tarea exige conocer la altura de la muralla. En un despliegue de observación científica, los sabios deciden estimar la altura de la muralla para calcular la altura de la torre. Saben que la mu-

ralla está construida con ladrillos de dimensiones iguales; por ello, solicitan que se envíe una pequeña partida de soldados para recopilar datos aproximados sobre la altura del bastión, contando, desde una prudente distancia, el número de ladrillos. Cumplida la misión, los soldados regresan para informar sobre sus observaciones. Pero surge un inconveniente: cada soldado da una cifra distinta del número de ladrillos que cree haber contado y muy pocos de ellos coinciden entre sí. Para resolver el problema, los sabios deciden tomar como indicador los datos que más se repiten en la visión de los soldados; esto es, convienen en usar una medida de tendencia central, que será analizada en esta obra, a la que se denomina la *Moda* (No; nada que ver con Christian Dior)

1
INTRODUCCIÓN

Estadística: Definición:
Disciplina que tiene por objeto la recopilación, sistematización, análisis e inferencias de los datos necesarios para tomar decisiones con cierta probabilidad de riesgo.

Para cumplir sus objetivos, a Estadística se divide en dos grandes grupos de estudio.

Estadística Descriptiva
Es la rama que obtiene los datos, los recopila y sistematiza para convertirlos en información útil y describir los rasgos característicos de un objeto de estudio.

Con ese objeto desarrolla y usa técnicas que están implícitos en los programas computarizados, tales como el SPSS.

Inferencia Estadística
Es la rama de la Estadística que utiliza la información sistematizada por la Estadística Descriptiva para inferir aspectos importantes de una población dada

La Inferencia Estadística cumple su tarea con cierto grado de probabilidad y recurre, tal como la Estadística Descriptiva, a la información que logra de las muestras.

Estadístico
Es el nombre genérico de cualquiera de las medidas utilizadas por la Estadística Descriptiva; por ejemplo, la media aritmética de los ingresos de una una empresa.

Variable
Es una magnitud que varía pero que puede ser medida, manipulada o controlada. Suele estar relacionada con otras variables y cambiar en concordancia

Las definiciones dadas son las que nos servirán para introducir el presente capítulo. Sin embargo, a medida que vayamos necesitando, tendremos nuevas definiciones.

Estoy seguro que este sistema evita que el participante quede apabullado por un número excesivo de definiciones sobre aspectos que aún no conoce.

SPSS
Es el más conocido, completo y útil de los programas computarizado de Estadística

En razón de que nuestra metodología se basa en el proceso "aprender-haciendo" es que vamos a iniciar el capítulo con un primer acercamiento al programa SPSS.

Abrir el SPSS

Hacemos click en el programa instalado ya en la computadora
De inmediato, aparece la Caja de Diálogo 1.1

Caja de Diálogo 1.1

Click en "cancelar" y aparece la pantalla 1.1, que es un segmento de la pantalla completa, con el objeto de que se observe el final de la misma

Pantalla 1.1

En la parte inferior izquierda del segmento de pantalla aparecen dos leyendas: Data View y *Variable View.*

Por defecto, el SPSS se inicia en el formato de Data View, tal como se ve en la Pantalla 1.1; allí anotaremos los datos que nos servirán para realizar nuestra tarea.

Luego veremos que, el modo *Variable View* nos sirve para dar nombre a las variables y estructurarlas. Antes de empezar, definiremos algunos conceptos previos.

La Población

La *Población* es el total de los elementos potencialmente observables; v.g. el número de familias que vive en una ciudad determinada.

El levantamiento de estos datos, los que cubren a todas las familias que viven en una ciudad, se realiza por medio del diseño y la ejecución de un censo.

En la disciplina estadística muy raras veces se usa la Población, pues es muy difícil llegar a ella; además, es cara; más bien, usamos la *Muestra.*

La Muestra

Es una parte de la Población, la que, utilizando las técnicas que aprenderemos en este curso, representa adecuadamente todas las características que tiene la Población.

El Tamaño de la Muestra

Es el número de elementos que conforman una muestra.
Hay varias maneras de lograr los datos que necesitamos en una muestra.

La Encuesta

Es la recopilación sistematizada de datos que logramos de una población determinada y que luego transformaremos en información útil.

La Estadística generalmente trabaja sobre la base de muestras que se logran mediante las encuestas.

Precisamente, la primera tarea que realizaremos para iniciar nuestro trabajo, será diseñar una encuesta con las variables codificadas.

Diseño de una encuesta

Para diseñar una muestra, debemos definir las variables que vamos a usar.

Vamos a suponer que deseamos saber algunas características de los empleados de una empresa que produce bienes para el mercado nacional.

Para estructurar y dar los nombres a nuestras variables, hacemos click en el modo *Variable View* que aparece al lado de *Data View* en la Pantalla 1.1

Al hacer click en *Variable View*, se nos presenta la Pantalla 1.2

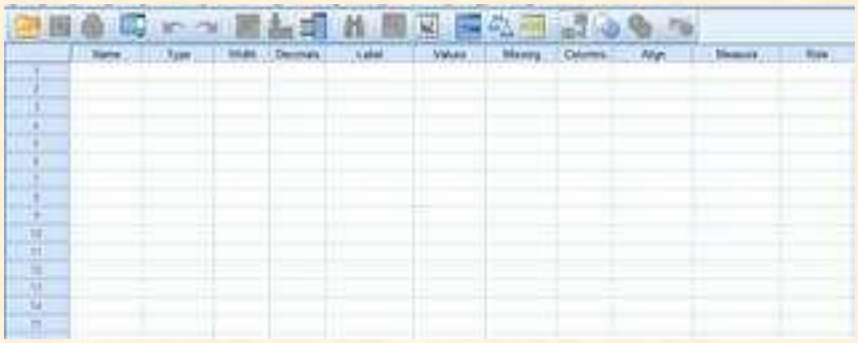
Supongamos que la primera primera variable que deseamos definir de las personas que serán encuestadas, se refiere a su género: hombre o mujer

En la fila No. 1 bajo la columna "Name", escribimos el nombre abreviado de género
De esta manera, el nombre de nuestra primera variable será "gene".

El ahorro de letras para nombrar a las variables es muy importante, como veremos después; por otra parte, hay una casilla especial para poner el nombre completo.

Obsérvese que a la derecha de "Name" está la casilla "Type".
Al pulsar en la casilla bajo la columna Type se nos aparece la leyenda "Numeric".

Pantalla 1.2



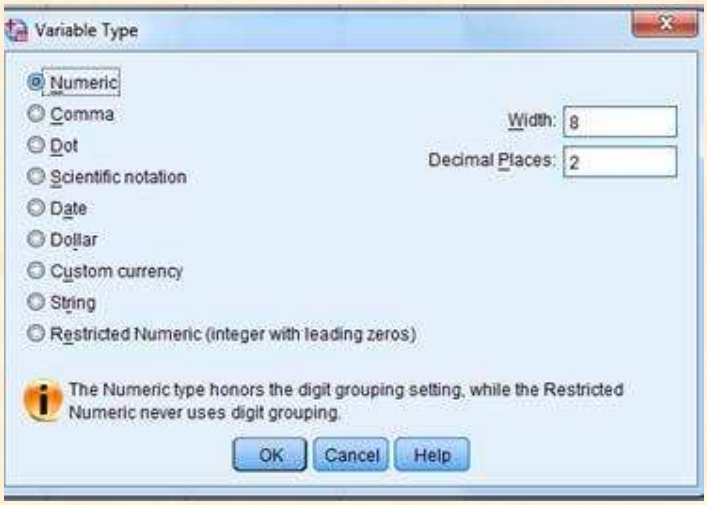
Click en esa casilla; de inmediato nos encontramos con tres puntos.
Click en esos tres puntos y tenemos la siguiente caja de opciones:

En la Caja 1.2 vemos una columna con varios nombres.
Cada uno de esos nombres establece la característica de nuestra variable

La variable "gene" es cualitativa (ya la vamos a definir) y debe ser clasificada como "string" en la lista de opciones del cuadro 1.1.

Sin embargo, para usarla apropiadamente, la convertidos en "Numeric" es decir, en una variable que pueda ser cuantificada

Caja de Diálogo 1.2



En la Caja 1.2 vemos una columna con nombres múltiples; cada uno de esos nombres describe las principales características de la variable que vamos a usar.

Esta es la razón de que en el menú de opciones de la Caja 1.2 escojamos "Numeric" Ahora, definamos algunos tipos de variables.

Variables Cuantitativas

Las conocemos como variables numéricas; este tipo de variables son las más comunes en los estudios estadísticos, pues varían en su magnitud.

Variables Categóricas

Son las variables cualitativas y se dividen, a su vez, en dos grandes ramas: las variables nominales y las variables ordinales.

Variables Nominales

Son aquéllas que no pueden ser clasificadas ni en una magnitud cuantitativa ni en una magnitud de jerarquía.

Por ejemplo, las categorías de género; *varón, mujer*, que es la variable que vamos a codificar, son variables de ese tipo.

Variables Ordinales

L*as* que aceptan una jerarquización de importancia.

El grado de Educación de las personas, por ejemplo, es una variable nominal, puesto que puede ser calificado de acuerdo a un orden, v.g, descendente.

La variable "genero" es una variable cualitativa-nominal, la que será codificada como númerica cuando necesitemos usarla.

La variable "gene" es una variable cualitativa y debería ser clasificada como "string" en la lista de opciones de la Caja 1.2

Sin embargo, para usarla apropiadamente, la convertimos en "Numeric", esto es, en una variable que puede ser cuantificable

Ésa es la razón por la que escogemos "Numeric" del menú de opciones de la Caja 1.2.

Codificación de la variable "género" para estructurar una muestra

Una vez que hicimos "clic" en la casilla "Numeric", pulsamos OK

La siguiente columna (witdh) nos pide establecer el ancho de la columna para la variable gene; hacemos clic en la casilla y aparece un menú de opciones

Pulsamos hasta 6, que será el ancho de nuestra columna

En la columna "Decimals" escogemos 0

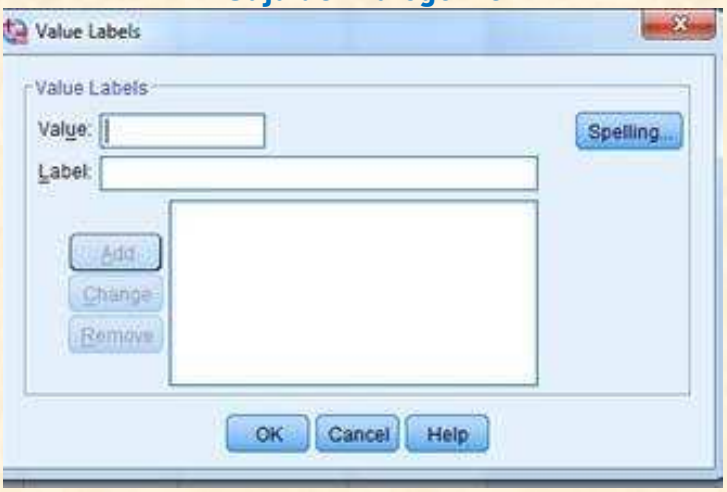
En la columna "Label" (etiqueta) ponemos el nombre formal de gene, en este caso, Género, pues ese será el nombre que constará en el informe final.

La siguiente columna se refiere a "Los Valores" que daremos a los géneros femenino y masculino; en realidad aquí es donde codificamos la variable nominal "Género"

Codificar significa dar a cada uno de los géneros un número que lo identifique
Para ello, hacemos clic en la casilla donde aparece la palabra "None"

Ahora nos encontramos con tres puntos a los cuales señalamos con el clic del mouse.
Inmediatamente aparecerá la Caja 1.3.

Caja de Diálogo 1.3



En la casilla, al lado de "Value" anotamos "1"
En la casilla al lado de "Label", anotamos: "Mujer"

Luego pulsamos Add y en la pantalla de abajo aparecerá "1 = Mujer"
La casilla al lado de Value aparecerá otra vez vacía.

Anotamos "2" y "Hombre", respectivamente y no olvidamos de pulsar Add
En la pantalla inferior aparecerá 2 = "Hombre"

La Caja 1.4 muestra el resultado del proceso.

Pulsamos OK y nos transportamos otra vez a la Pantalla Variable View
Allí vemos cómo quedó estructurada nuestra primera variable: *gene.*

Dejamos para más adelante las demás columnas.
La Pantalla 1.4 (Variable View) muestra lo que hemos conseguido hasta ahora.

Caja de Díalogo 1.4

Ahora regresamos a "Data View", haciendo click en la casilla inferior izquierda de la pantalla para ver cómo nuestra primera variable, *gene*, aparece en la primera columna

Hemos estructurado nuestra primera variable, después de convertirla en numérica
En cada columna de Pantalla 1.4 se muestra cada característica de la variable.

Supongamos que la segunda variable es la *edad.*

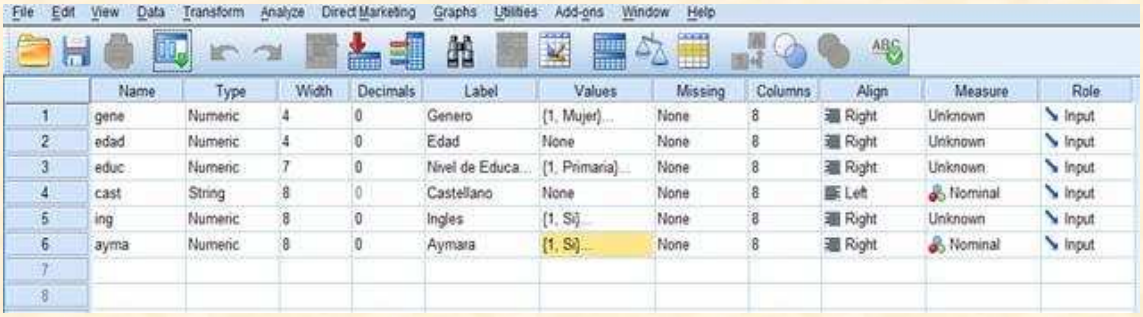
Con esta variable no hay problema, pues es cuantitativa y no necesita se codificada.
En la pantalla Variable View anotamos las características de la variable

Nos interesa el nivel de educación, que es una variable Categórica ordinal
Pero, al igual que *gene* la convertimos en Numérica.

Hasta el momento, hemos codificado las variables en el caso de que fueran categóricas, ya se nominales u ordinales, pero que debíamos escoger una de ellas

Por ejemplo, la variable idio, que es ordinal, no tiene esa limitación.
Una persona puede hablar más de un idioma

Pantalla 1.3

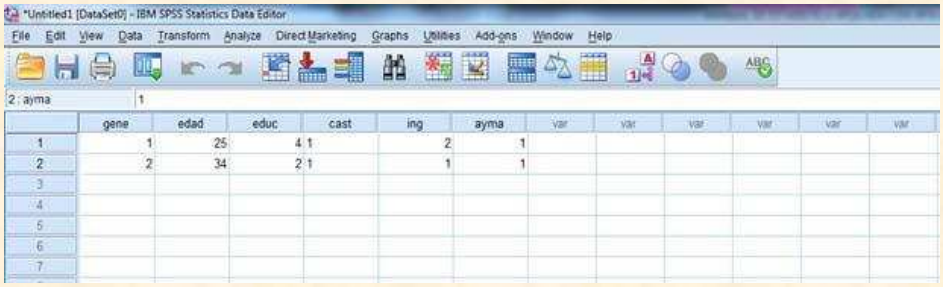


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	gene	Numeric	4	0	Genero	{1, Mujer}...	None	8	Right	Unknown	Input
2	edad	Numeric	4	0	Edad	None	None	8	Right	Unknown	Input
3	educ	Numeric	7	0	Nivel de Educa...	{1, Primaria}...	None	8	Right	Unknown	Input
4	cast	String	8	0	Castellano	None	None	8	Left	Nominal	Input
5	ing	Numeric	8	0	Ingles	{1, Si}...	None	8	Right	Unknown	Input
6	ayma	Numeric	8	0	Aymara	{1, Si}...	None	8	Right	Nominal	Input
7											
8											

Por lo tanto, registramos en la pantalla Variable View cada idioma por separado.
La Pantalla Variable View queda estructurada como se muestra en Pantalla 1.3

La Pantalla de Data View queda como se muestra en la pantalla 1.4, a la que hemos puesto valores en cada casilla, simulando una mini-encuesta

Pantalla 1.4


*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
2 : ayma
gene edad educ cast ing ayma var var var var var var
1 1 25 4 1 2 1
2 2 34 2 1 1 1
3
4
5
6
7

De esta manera, hemos estructurado el formulario de nuestra primera encuesta

2
MEDIDAS DE TENDENCIA CENTRAL

Una Medida de Tendencia Central es el punto medio de una distribución de datos.

Ahora estudiaremos las características de las más importantes medidas con breves ejemplos manuales y también haciendo uso del SPSS.

La Media Aritmética

La conocemos como el promedio de una serie de datos.

Supongamos que deseamos saber el promedio de la edad de 6 personas que están en un consultorio médico, haciendo un tratamiento especial.

Para conocer ese promedio, sumamos las edades de todos y cada uno; luego dividimos el resultado entre 6, que es el número de personas que nos interesa.

Media aritmética = (40 + 34 +28 + 10 + 23 + 50)/6 = 30.83
La media aritmética o promedio de las edades de las seis personas es 30.83 años.

Si queremos saber el ingreso promedio de 5 amigos para ir a tomar un café, preguntamos a cada uno cuanto tiene y luego hacemos lo mismo que en el caso anterior

Los datos, hipotéticos, que obtenemos de los cinco son los siguientes:

El primer amigo tiene $4
El segundo, $20

El tercero, $12
El cuarto, $9
El quinto, $13

Media aritmética = (4 + 20 + 12 + 9 + 13)/5 = 11,60

El resultado nos hará saber que, en promedio, se tiene 11 dólares con 60 centavos, independientemente de la suma que cada uno tenga en la realidad.

Ese es el concepto fundamental de la media aritmética.
Luego veremos porqué es tan útil.

No importa cuán pequeña o grande pueda ser el conjunto de datos que se nos presente, la media aritmética siempre tendrá la misma definición:

La suma de todos los valores, dividida entre el número de casos.
Usando símbolos, tendremos:

$$X^* = \Sigma Xi/n$$

X* es la media aritmética de la muestra que hemos tomado; Xi, representa a cada uno de los valores que serán sumados; n es el número de observaciones.

La misteriosa Σ indica que todos los valores Xi deben ser sumados.

Empezando con el SPSS

Definida conceptualmente la Media Aritmética vamos al SPSS

Para realizar estos ejercicios, el SPSS cuenta con una lista de Samples, esto es, datos ya registrados que nos ayudan a realizar los ejercicios requeridos

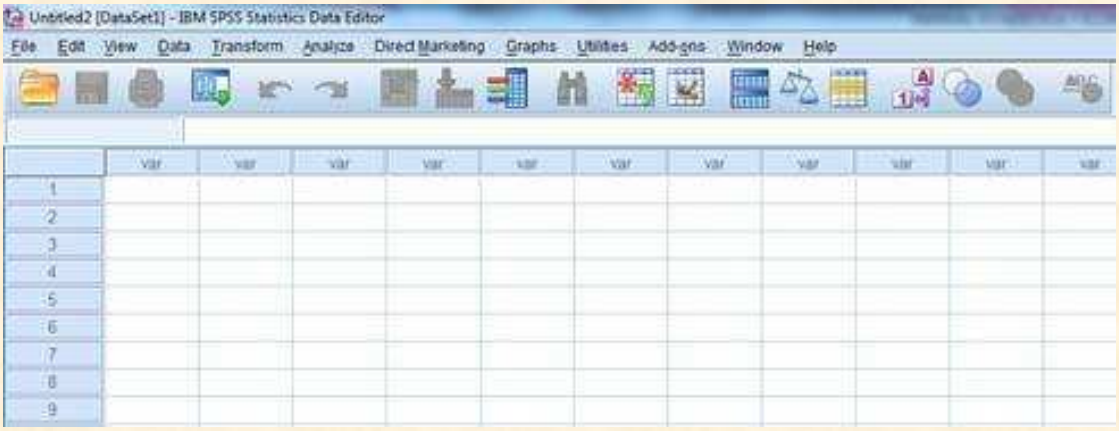
Supongamos que deseamos saber la media aritmética de los sueldos actuales que reciben los empleados de una empresa determinada.

Supongamos que deseamos saber la media aritmética de los sueldos actuales que reciben los empleados de una empresa determinada.

En el paquete del SPSS hay una gran base de datos que viene con el programa

Para mostrar cómo llegar llegar a esos archivos, traemos a esta página un segmento de la pantalla de SPSS.

Pantalla 2.1


Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Con el mouse vamos al menú principal de la Pantalla 2.1 y ejecutamos los siguientes comandos en el orden que se anota a continuación

Click en Menú → File → Open → Data

*Se nos presentará la Caja 2.1; a*hí están todos los archivos que trae el SPSS bajo el título de Samples. Esos archivos están registrados por orden de abecedario

Caja de Diálogo 2.1

Los archivos son usados para que realicemos los ejemplos necesarios en cualquier tema que convoque alguna función del menú

Ahora necesitamos encontrar el archivo *Employee data.sav*
La extensión *.sav nos indica que el archivo pertenece al SPSS

Encontrado el archivo, Click *Open* y tendremos la Pantalla 2.2 del SPSS en la modalidad Data View, de la cual traemos un fragmento para analizar los datos.

Pantalla 2.2

Employee data.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0	
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0	
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0	
4	4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0	
5	5	m	02/09/1955	15	1	$45,000	$21,000	98	138	0	
6	6	m	08/22/1958	15	1	$32,100	$13,500	98	67	0	
7	7	m	04/26/1956	15	1	$36,000	$18,750	98	114	0	
8	8	f	05/06/1966	12	1	$21,900	$9,750	98	0	0	
9	9	f	01/23/1946	15	1	$27,900	$12,750	98	115	0	

Notamos que en la esquina superior derecha está el nombre del archivo que estamos usando: *Employee data.sav*

Nuestra tarea consiste en averiguar la media aritmética de los salarios actuales que reciben los trabajadores de la empresa

Con el curso apuntamos a la casilla Salary por algunos segundos sin hacer ckick
Aparecerá una leyenda complementaria: Current Salary (Salarios actuales)

Current S*alary es el nombre formal de salary*
Sin embargo, de entrada vemos que hay un problema.

Los salarios están registrados con el signo $ = dólar
Para calcular la media aritmética tenemos que cambiar los datos al tipo *Numérico*

Para el efecto, realizamos la misma operación que hicimos en el capítulo anterior
Vamos a repetir el procedimiento

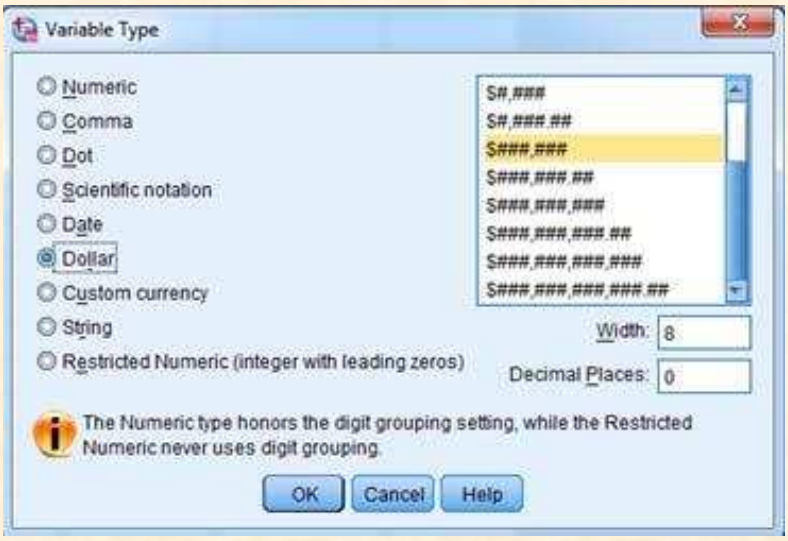
De la modalidad *Data View* pasamos a la de *Variable View*
Allí están registradas todas las características de la variable salary

En la columna Type, observamos que nuestra variable, salary, está inscrita en dólares
Para cambiar a la función Numeric usamos los siguientes comandos

Click → Dollar → click en los tres puntos…

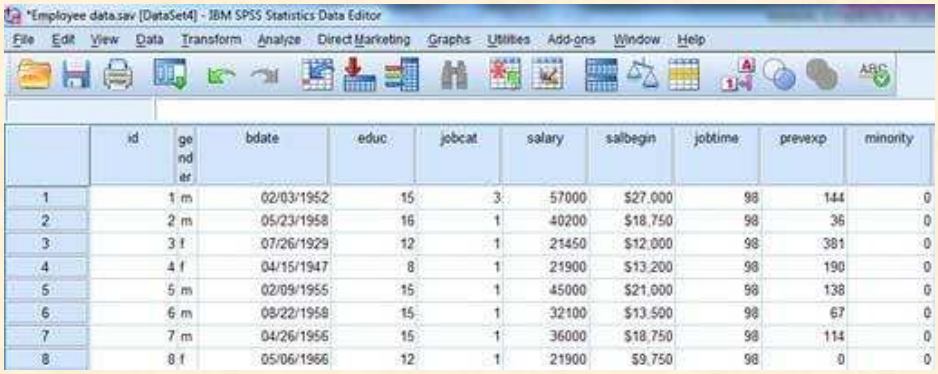
La Caja 2.2 nos muestra que nuestra variable está registrada en términos de dólar

Caja de Diálogo 2.2



Click → Numeric → OK
El SPSS nos trae la pantalla 2.3:

Pantalla 2.3


*Employee data.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority
1	1	m	02/03/1952	15	3	57000	$27,000	98	144	0
2	2	m	05/23/1958	16	1	40200	$18,750	98	36	0
3	3	f	07/26/1929	12	1	21450	$12,000	98	381	0
4	4	f	04/15/1947	8	1	21900	$13,200	98	190	0
5	5	m	02/09/1955	15	1	45000	$21,000	98	138	0
6	6	m	08/22/1958	15	1	32100	$13,500	98	67	0
7	7	m	04/26/1956	15	1	36000	$18,750	98	114	0
8	8	f	05/06/1966	12	1	21900	$9,750	98	0	0

Allí vemos que el signo $ = Dólar ha desaparecido

Ahora sí, estamos listos para averiguar la Media Aritmética de los salarios mensuales que reciben los trabajadores, empleados y ejecutivos de la empresa

Menú → *Analize*.

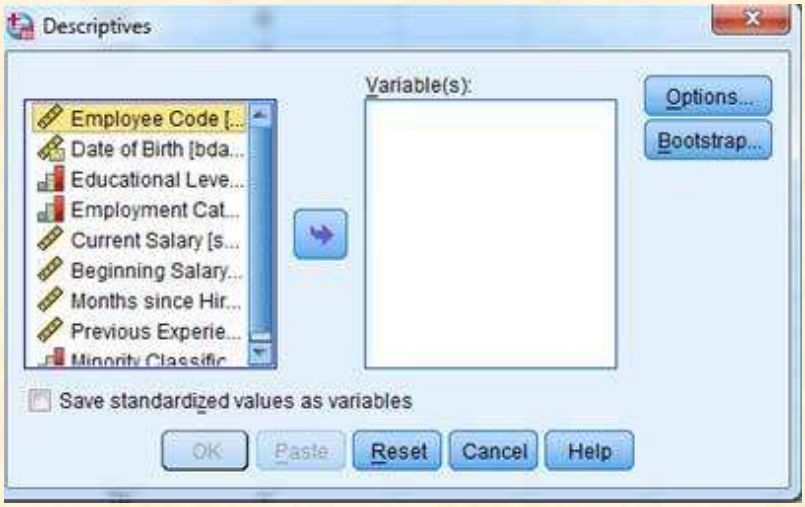
Se nos presentará un menú vertical de opciones; pulsamos *Estadísticos Descriptivos*
A la derecha aparecerá otro cuadro, del cual escogemos *Descriptivos*

De inmediato nos damos cuenta que las variables están registradas con sus nombres formales; también vemos un cuadro en blanco a la derecha.

Pulsamos *Current salary* en el cuadro de la izquierda y también en la flecha que apunta a la pantalla en blanco de la derecha; *Current Salary* se trasladará a esa pantalla

Se nos presentará la Caja 2.3 con las variables de *Employee data.sav* a la izquierda.

Caja de Diálogo 2.3



Click en Options y el SPSS nos muestra el cuadro pequeño dentro de la pantalla 2.4

Pantalla 2.4

Hay varias opciones; por el momento sólo nos interesa la Media, por lo tanto pulsamos todos los botones para borrarlas y nos quedamos sólo con el botón respectivo

Mean → Continue → OK

El SPSS nos lleva a una pantalla de Resultados

Es en esa pantalla donde el SPSS anota los resultados de los trabajos que realizamos con los datos del modo *Data View*

En esa Pantalla de Resultados se consigna la Tabla de Resultados 2.1

Tabla de Resultados 2.1

	N	Mean
Current Salary	474	34419,57
Valid N (listwise)	474	

En la figura Cuadro de Resultados 2.1, la columna bajo la letra N señala el número de trabajadores de la empresa, 474

La columna etiquetada con Mean nos da el valor de la Media Aritmética
Mean = $34.419,57

Para borrar la pantalla de resultados, dirigimos el Mouse a la esquina superior izquierda, hacemos click en la casilla Output y borramos con la tecla de supresión del teclado.

Si pulsamos las opciones Minimun y Maximun Continue → OK tendremos la Tabla de resultados 2.2

Tabla de Resultados 2.2

	N	Minimum	Maximum	Mean
Current Salary	**474**	**15750**	**135000**	**34419,57**
Valid N (listwise)	**474**			

La columna Mínimun muestra que el salario más bajo registrado es 15.750.
El dato bajo la columna Maximun nos dice que el salario más alto es 135.000.

Hay muchas más opciones para complementar la información que viene con la media aritmética, las que usaremos después

Por el momento, debo decir que Nuestro primer objetivo está cumplido:

Hemos logrado calcular, usando el SPSS, la media aritmética de los sueldos y salarios que reciben los 474 empleados que trabajan en la empresa.

Pero, en el proceso hemos cumplido otra tarea importante: aprender algunos comandos básicos del SPSS

Como un ejercicio muy útil pueden estimar la media aritmética de los salarios que tenía cada trabajador el día que ingresó a la empresa.

Esos datos están en la columna salbegin del menú de la pantalla

La Media Aritmética es la medida de tendencia central más conocida y usada que todas las demás, tiene un punto débil que debemos tomar en cuenta

Es muy susceptible a variar con los valore extremos
Para constatarlo, usemos como ejemplo el ejercicio manual referido a los ingresos

Dijimos que la media aritmética de los ingresos de cinco amigos era:

$$(4 + 20 + 12 + 9 + 13)/5 = 11,60$$

Ahora supongamos que aparece uno más y se suma a la propuesta de tomar café
Asumamos que su ingreso es $100

Sumamos la nueva serie y la dividimos entre 6

$$(4 + 20 + 12 + 9 + 13 + 100)/6 = 26.34$$

La nueva media se ha incrementado en más del doble por la inclusión de un valor alto

Cuando creemos que hay valores extremos como el caso anterior ya no usamos la media aritmética, sino que recurrimos a otra medida de tendencia central: la Mediana

La Mediana

Es el valor que está más al centro de un conjunto de datos ordenados
Para consolidar el concepto, realicemos primero un ejercicio manual

120, 140, 200, 240, 260, 380,450, 500, 630, 700, 750

Observemos que los datos están ordenados de menor a mayor

En este caso, la Mediana es $380, pues este valor hace que el conjunto de observaciones a su izquierda (5) iguale al conjunto de observaciones a su derecha (5)

Pero, al igual que la Media Aritmética, si queremos estimar la Mediana de 5000 observaciones, el trabajo se hace pesado

Sobre todo si tenemos que expresar los valores en forma ascendente, desde el más pequeño al mayor

Para evitarnos esa descomunal tarea, convocamos a nuestro amigo SPSS
Vamos a realizar el ejercicio con el SPSS sobre el archivo *Employee data.sav*

Utilizamos los mismos comandos que en el ejemplo de los salaries

Menú → Open → Data → Employee data.sav → Open

Una vez que tenemos el archive abierto pulsamos la siguiente serie de comandos:

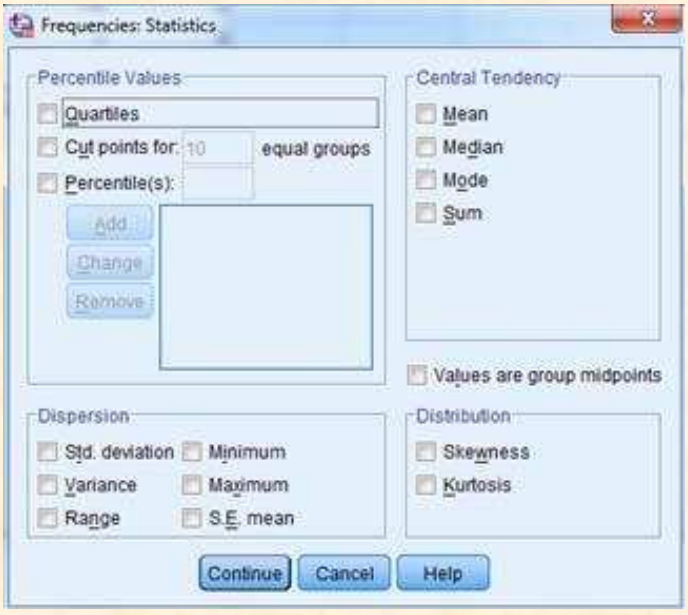
Menu → Analize → Descriptive Statistics → Frecuencies

La Caja que inscribe la lista de las variables del archivo que vamos a utilizar aparece en la pantalla

Click en *Current Salary* en la flecha de dirección
La variable *Current Salary*, se trasladará a la pantalla de la derecha

En el nuevo cartel de opciones presionamos el botón *Statistics*
Tendremos la Caja 2.4; en la parte derecha hay un menú de opciones

Caja de Diálogo 2.4



Median → Continue → OK

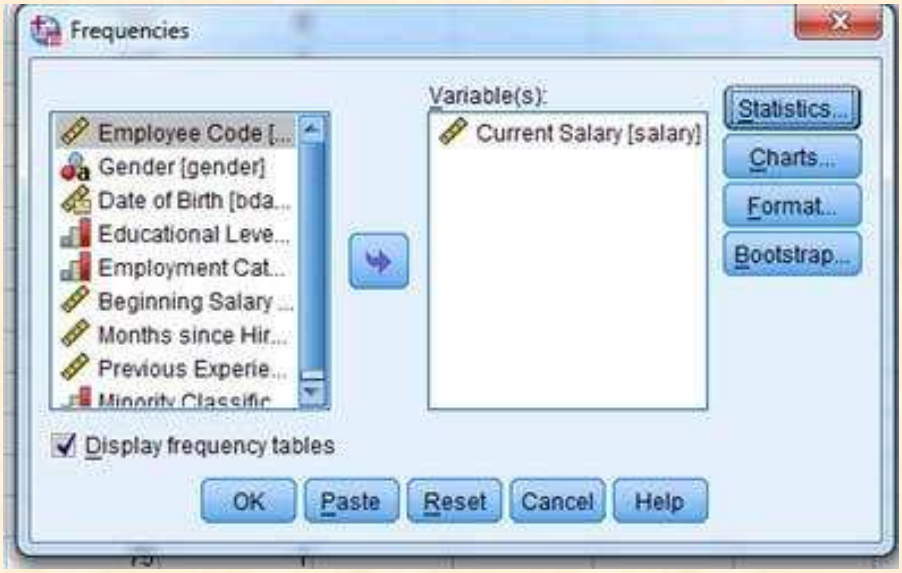
La Pantalla de Resultados nos muestra algo que no teníamos en mente, esto es, todas las observaciones que estructuran el archivo, algo que no necesitamos

Para solucionar el problema borramos la pantalla de resultados y volvemos a ejecutar los primeros comandos, tal como sigue:

Menu → *Analize* → *Descriptive Statistics* → *Frecuencies* → *Statistics* → *Median* → *continue*
En este punto aparece la Caja 2.5

Observemos que en la parte inferior izquierda el botón: Display frequency tables está habilitada; esa opción es la que nos trajo la lista de todos los trabajadores

Caja de diálogo 2.5



Hacemos clik en ese boton para eliminar la opción respectiva y luego presionamos OK
La Tabla de Resultados ha diseñado la Tabla de Resultados 2.3

Tabla de Resultados 2.3

N	**Valid**	**474**
	Missing	**0**
Median		**$28,875**

La Mediana de los salarios del personal de la empresa es $28,875
Este indicador es menor que el valor de la Media Aritmética: 34419,57

La diferencia se debe a que la Media Aritmética estaba sesgada hacia los valores muy altos y no reflejaba la verdadera situación de los promedios salariales

Pero el repertorio de las medidas de tendencia central no acaba aquí, pues tenemos otra de igual importancia que las demás.

La Moda

Es el valor que más se repite en el conjunto de datos y que los griegos que deseaban fabricar una torre para desbaratar la ciudad sitiada utilizaron como indicador.

Tomemos los datos utilizados en la mediana, pero lo transformemos de tal manera que el valor 260 se repite tres veces; en ese caso la Moda será 260.

120, 140, 200, 240, 260, 260, 260, 380, 390, 450, 630, 700, 750, 780

La Moda se aplica en el caso de que haya varios valores repetidos en la muestra.
Para disponer de la Moda, seguimos la siguiente serie de comandos ya conocidos

Menu → Analize → Descriptive Statistics → Frecuencies → Statistics → Mode → Continue

Tendremos en pantalla la Tabla de Resultados 2.4 en la que en vez del botón Median, pulsamos Mode → OK

Tabla de Reultados 2.4

N	**Valid**	**474**
	Missing	**0**
Mode		**$30,750**

La Moda es $30,750, valor que es el que más se repite en el archivo *Employee data. save*
En este caso su valor es mayor a la Mediana, pero menor que la Media

Cuando lleguemos al capítulo de las distribuciones, usaremos una gráfica de distribución para comparar la media aritmética, la mediana y la moda, entre sí.

Mientras tanto, diremos que, de todas las medidas de tendencia Central, la Media Aritmética es la más utilizada.

Tales son las medidas de tendencia central más importantes.
Ahora analizaremos las medidas de dispersión

Medidas de Dispersión

Vimos que las medidas de tendencia central identifican un valor que se acerca más al centro de una serie de datos o de elementos.

Las medidas de dispersión nos muestran el grado en que se alejan del centro.
La primera medida de dispersión, la más simple, es el rango.

El Rango

Es la diferencia entre el valor más alto y el más pequeño de los datos.
Acudamos otra vez a las serie de datos utilizados para calcular la Mediana:

120, 140, 200, 240, 260, 260, 380, 390, 450, 630, 700, 750, 780

El Rango será 780 – 120 = 660

Al igual que para el cálculo de la Mediana, el SPSS ordenará cualquier conjunto de datos que no estén ordenados.

El Rango nos da una primera percepción sobre las diferencia extremas que median en una serie de datos, en este caso, en un registro de salarios.

Medidas de desviación promedio

Calcula la desviación promedio entre los valores de una serie de datos y una medida de tendencia central; primero nos interesa el concepto.

La Varianza

Es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística; la varianza de la muestra se representa por S^2

Para aclarar el concepto, tomemos los siguientes datos: 2, 4, 6, 8, 10
La media aritmética será: (2 + 4 + 6 + 8 + 10)/5 = 30/5 = 6

Con ese dato, analicemos los valores originales.

El valor 2 se desvía en - 4 de la media aritmética (2 - 6 = - 4)
El valor 4 se desvía en - 2 unidades: (4 - 6 = - 2)

El valor 6 no se desvía de la media aritmética: (6 – 6 = 0)
El valor 8 sí se desvía en 2 (8 – 6 = 2) que es una desviación positiva.

Lo mismo sucede con el valor de 10, su desviación positiva es 4 (10 – 4) = 6

Si deseáramos estimar la media aritmética de estas desviaciones nos encontraríamos que su valor sería 0, pues los valores positivos anularían a los negativos.

Para eliminar este problema no tomamos en cuenta la desviación simple de cada observación con la media aritmética; más bien elevamos cada desviación al cuadrado

Lo hacemos así, porque una cantidad elevada el cuadrado, ya sea positiva o negativa, siempre nos dará un resultado positivo, que es lo que se busca.

La suma de los cuadrados de esas diferencias será dividida por el número de observaciones, que es 5, al que le restaremos 1.

Así lo determinaron los grandes matemáticos.
El denominador será 5 – 1

$$S^2 = [(2 - 6)^2 + (4 - 6)^2 (6 - 6)^2 + (8 - 6)^2 + (10 - 6)^{2]}]/(5 - 1)$$
$$S^2 = [(-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2]4 = [(16 + 4 + 4 + 16)]/4 = 40/4 = 10$$

En consecuencia diremos que la Varianza de la Muestra es 10

Cuando el número de observaciones es alto, el cálculo de la varianza se hace muy pesado por lo que recurrimos, como de costumbre, a nuestro amigo el SPSS

Volvamos a nuestro file *Employee data.save*

Analize → Descriptive Statistics → Descriptives → Currente Salary → Options

En la Caja que aparece, puede haber varios botones que esén marcados
Los desemarcamos y dejamos solo → *Variance → Continue → OK*

En la tabla de resultados tendremos la Tabla de Resultados 2.5

Tabla de Resultados 2.5

	N	Variance
Current Salary	**474**	**291578214**
Valid N (listwise)	**474**	

La Desviación Típica de la Muestra = s
Es la raíz cuadrada de la Varianza: $s = \sqrt{S^2}$

En este caso, s = raíz cuadrada de 291578214,453 = $\sqrt{291578214}$ = **17075**
Pero vayamos al SPSS para estimarla debidamente

Analize → Descriptive Statistics → Descriptives → Currente Salary → Options → Std.deviation → Continue → OK

Tabla de Resultados 2.6

	N	Std. Deviation
Current Salary	474	$17,075
Valid N (listwise)	474	

Confirmamos que la desviación estándar de los salarios es, efectivamente, $17,075

El Coeficiente de Variación

Es la relación entre la desviación típica y la media de la muestra: s/X*

En el desarrollo de los indicadores estadísticos de *Current Salary*, vimos que la media aritmética es 34419 y la desviación estándar es 17,075

*Coeficiente de Variación = s/X** = $\frac{17075}{34419}$ = *0.50*

Las medidas de tendencia central y de dispersión que hemos analizado son los indicadres básicos de la *Estadística Descriptiva.*

Más adelante estableceremos la utilidad de estos indicadores.
Hasta ahora hemos estimado cada indicador por separado

Lo hicimos así para que logremos una adecuada conceptualización de cada una

Sin embargo, en los problemas reales no estamos interesados sólo en un indicador, sino que requerimos un conjunto de todos ellos

Haremos un ejercicio integral para obtener todos los indicadores que hemos visto hasta ahora sobre la variable Current Salary" del archivo *Employee data.sav*

Analize → Descriptive Statistics → Frequencies → Currente Salary →

→(borrar display frequency tables) → Statistics → Options →Mean → Median → Mode →Minimum →Maximun → Range → Std. deviation → Variance → Continue → OK

Tabla de Resultados 2.7

N	Valid	474
	Missing	0
Mean		$34,419.57
Median		$28,875.00
Mode		$30,750
Std. Deviation		$17,075.661
Variance		291578214,453
Range		$119,250
Minimum		$15,750
Maximum		$135,000

Con este ejercicio concluimos el capítulo referido a los indicadores más importantes de la Estadística Descriptiva.

3
DISTRIBUCIÓN DE FRECUENCIAS

Concepto

Una distribución de frecuencias es una tabla en la que organizamos los datos dividiéndolos en Clases o grupos que describen alguna característica de la población.

Si no dividimos los datos en grupos, cuando pidamos, v.g, las frecuencias de los salarios, el SPSS nos dará un cuadro con el salario de cada uno de los empleados.

Si la fábrica tiene 2000 empleados, tendremos un cuadro de 2000 salarios.

Sin embargo, cuando necesitamos las medidas de tendencia central o cualquier otra, el SPSS acudirá a los datos originales, no a los grupos.

A no ser que, por alguna razón, necesitemos la media u otro estadístico de cada uno de los grupos.

Una distribución de frecuencias muestra el número de observaciones del conjunto de datos que caen en cada una de las clases en las que hemos dividido los datos.

Utilizaremos el SPSS y el archivo *Employee data.sav* para construir el cuadro de frecuencias del salario actual (*Current Salary*) de los empleados de la empresa.

Menú Principal → Transformar → Recodificar en distintas variables.

Abierto ya el cuadro de diálogo vemos dos pantallas.

La pantalla de la izquierda contiene todas las variables; escogemos *Current Salary* y la trasladamos a la pantalla de la derecha, haciendo click en la flecha de dirección.

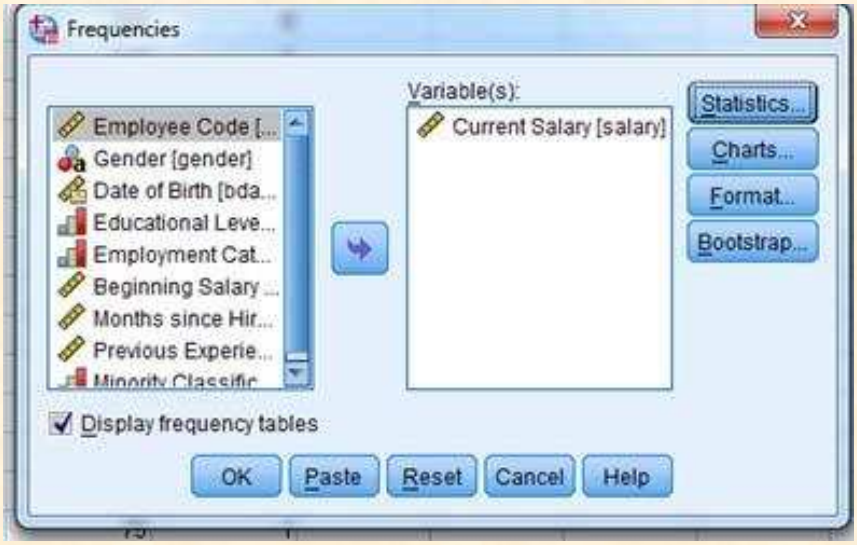
La casilla *Current Salary* ha cambiado a *salary*, que es el nombre que aparece al comienzo de la columna respectiva, mientras que *Current Salary*, es la etiqueta formal

En el cuadro aparece un signo de interrogación después de *salary*, eso significa que el SPSS nos pide dar un nuevo nombre a la variable que vamos a recodificar.

En la parte derecha hay dos casillas: una que dice *nombre* y la otra *etiqueta.*

Ingresamos a la casilla "*nombre"* y registramos el nombre que deseamos ponerle a la nueva variable, digamos *rsalay* (Podría ser cualquier otro) Cuadro 3.1

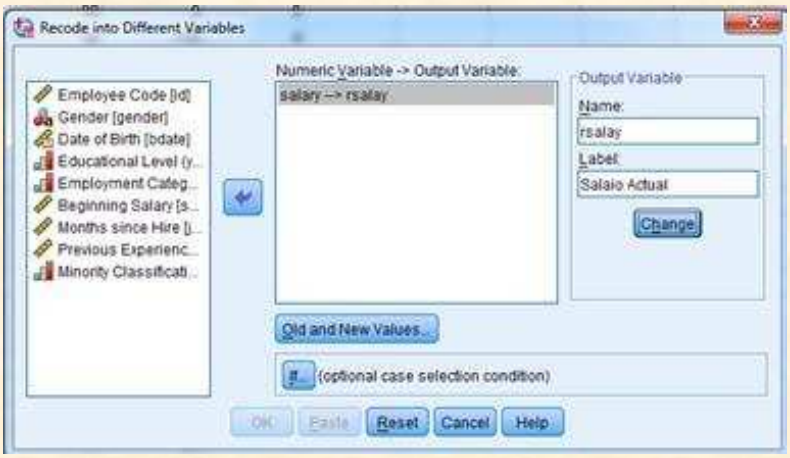
Caja de Diálogo 3.1



En la casilla "Etiqueta" ponemos el nombre formal *Salario Actual*; pulsamos la casilla que está debajo de Label, *Cambiar*

La variable *rsalay* ocupa el lugar donde estaba el signo de interrogación, tal como se muestra en el Cuadro 3.2.

Caja de Diálogo 3.2

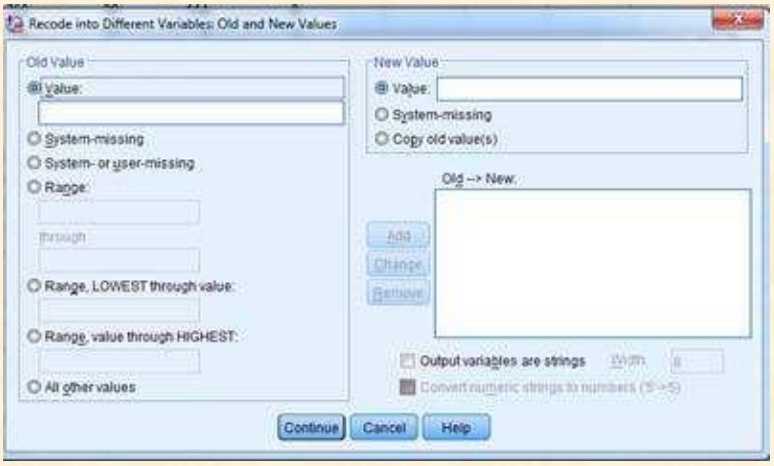


Así, hemos estructurado una nueva variable, *rsalay*, sobre la base de la variable *salary; d*esde este momento, el SPSS tratará a la nueva variable como independiente.

Ahora pulsamos en la casilla de la parte inferior: *Valores antiguos y nuevos debajo de la pantalla blanca, y* tenemos un cuadro de diálogo

El cuadro 3.3 es el que nos ayudará a dividir nuestros datos en clases.

Caja de Diálogo 3.3



Lo que vamos a hacer es codificar las clases, como codificamos la variable *sexo.*

Clase

Es un intervalo con valores mínimos y máximos en que dividimos los datos para analizar más apropiadamente sus indicadores.

El número de clases en que se divide el total de los elementos de un archivo, depende del volumen de datos y de la percepción del analista

Sin embargo es necesario tener una idea de los valores mínimos y máximos de la variable cuyos elementos serán clasificados en grupos o clases

En este caso, dividiremos los datos en cuatro clases

Para ello pulsamos el botón que dice: *rango*, en la parte izquierda del cuadro
Al hacerlo, se habilitarán dos casillas en blanco.

En la primera registramos 15000; en la casilla de abajo, escribimos 45000.
Así tendremos el número de empleados que ganan entre 15000 y 45000 al año

Vamos a la parte superior derecha click → *valor* en la casilla anotamos 1.

Click en *Add* y en la casilla inferior aparece 15000 *trhu* 45000 → 1.
La primera de nuestras clases incluirán todos los valores desde 15000 hasta 45000.

Otra vez vamos a la izquierda click → "rango" y anotamos 45001 en la casilla de abajo; vamos a la casilla inferior, donde dice "hasta" y anotamos 75000.

Vamos al lado derecho → "Valor", anotamos 2 en la casilla respectiva → Add

En la pantalla inferior aparecerá otro registro después del anterior.
Esta vez con la leyenda que dice: *45001 thru 75000 → 2.*

Para la tercera clase; anotamos 75001 en la casilla *rango* y 105000 en la que dice *thru*

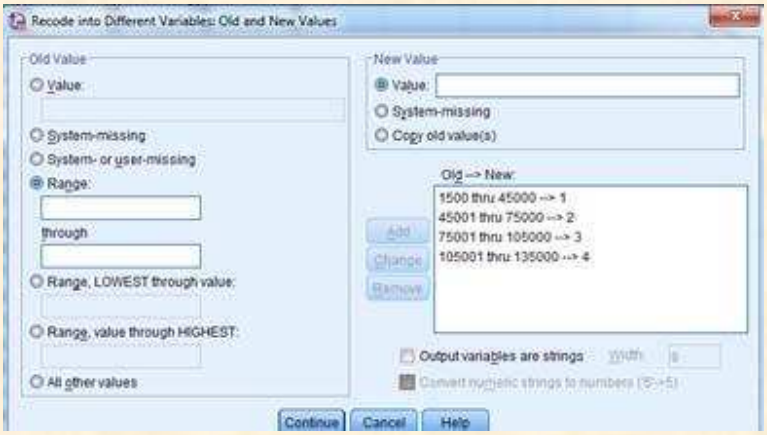
Al lado derecho →valor → 3 → Add

Por último, hacemos lo mismo para la cuarta: 105001 thru 135000

Al lado derecho valor → 4

Ahora tenemos los datos completos tal como aparecen en el cuadro 3.4

Caja de Diálogo 3.4



Continue → OK

El SPSS nos lleva a la pantalla de resultados, pero los resultados no están allí
Están en la última columna de *Vista de Datos* con el nombre rsala

En la última columna de aparecerán los intervalos codificados 1, 2, 3 y 4, de acuerdo con los valores de cada rango salarial.

Gráficas de las distribuciones de frecuencias: El Histograma

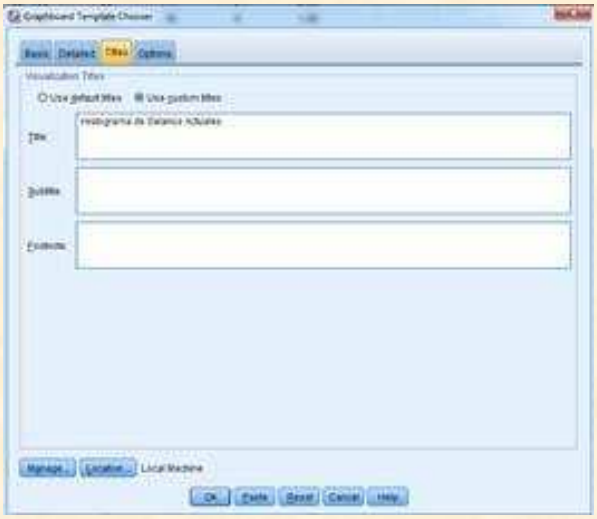
Es una gráfica que consiste en una serie de rectángulos, el ancho de cada uno mide la distancia que existe entre las cantidades que estructuran una clase estratificada.

La distancia vertical nos da los valores para esa clase.
Para obtener el Histograma respectivo procedemos del siguiente modo:

Menú Principal → Gráficas → Legacy Dialog → Histograma
Pulsamos en Histograma → Títulos

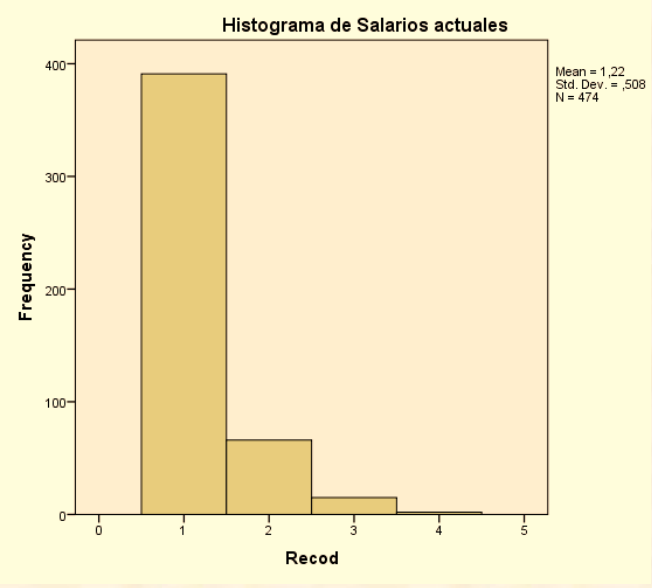
Escribimos el título Histograma de Salarios Actuales→ OK

Caja de Diálogo 3.5



En la Pantalla de Resultados del SPSS veremos la Gráfica 3.1

Gráfica 3.1



El Histograma es de gran ayuda, especialmente cuando queremos comprobar si la distribución de una variable se aproxima a la normal, tal como veremos después.

Es un diagrama importante para los técnicos en estadística porque las clases están representadas por números, los que no dicen mucho a los ejecutivos de la empresa

Por todo lo expuesto, deseamos es que en la *Pantalla de Datos* aparezcan, textualmente, los intervalos, tales como 15000 – 45000 para todos los rangos que hemos recodificado.

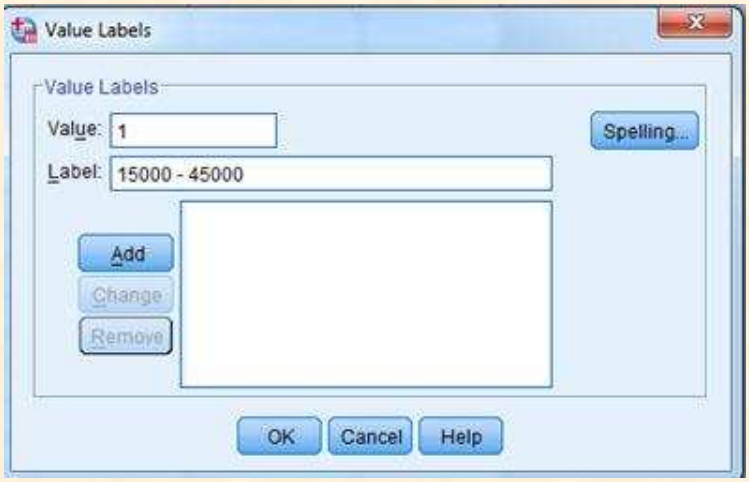
Vamos a *Vista de Variables,* la nueva variable, *rsalay está en la última fila*
En la columna *Value*, aparece el rótulo "Ninguna".

Click en los puntos y se nos abre el cuadro de diálogo que ya conocemos.

En la casilla *Valor* anotamos 1 y en la que dice Etiqueta registramos 15000 – 45000, pulsamos *Añadir* y esta primera clase aparece registrada en la pantalla

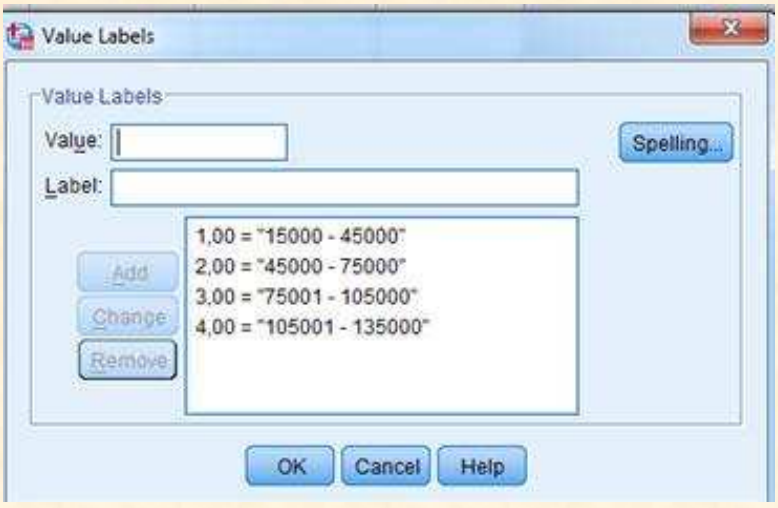
Pulsamos Add y la clase 15000 – 45000 aparecerá en la pantalla inferior

Caja de Diálogo 3.6



Realizamos el mismo ejercicio con las demás clases

Caja de Diálogo 3.7



Pulsamos OK y el SPSS nos lleva a la Pantalla de Datos
Para que las clases se vean, → *Menu* → *Value Levels*

En la última columna tenemos ya las clases debidamente codificados

La primera clase 15000 – 45000 fue registrada sin ninguna variación, pero la siguiente empieza con 45001, para no repetir el valor de 45000; lo mismo con las demás.

Ahora nos corresponde saber cuántas observaciones hay en cada clase.

Frecuencias

En la pantalla de *Vista de datos*, vamos al menú superior

Analizar; → *Descriptive Statistics* → *Frequencies*

Se abre el cuadro en el que la lista de variables está a la izquierda.
La nueva variable que buscamos está al final con el nombre *Salario actual (rsalay)*

C*lick* → *Salario actual (rsalay)* y con la flecha de dirección la llevamos a la derecha

Notamos que en la parte inferior izquierda el botón Display Frequency tables" está habilitada, ahora sí lo necesitamos de ese modo

Aceptamos.

En la Pantalla de Resultados aparece el cuadro de frecuencias con las clases respecti vas; es el que nos servirá para diseñar nuestra primera distribución de frecuencias.

Tabla de Resultados 3.1

		Frequency	Per-cent	Valid Per-cent	Cumulative Percent
Valid	15000 - 45000	391	82,5	82,5	82,5
	45000 - 75000	66	13,9	13,9	96,4
	75001 - 105000	15	3,2	3,2	99,6
	105001 - 135000	2	,4	,4	100,0
	Total	474	100,0	100,0	

En la Tabla de Resultados 3.1 observamos que hay 391 empleados que ganan entre 15000 y 45000; el porcentaje de esos empleados (*Valid Percent*) es del 82%

La misma interpretación tendremos para las demás clases.

Como podemos observar, los salarios de los 474 empleados se han agrupado en cuatro clases, fáciles de entender en vez de tener una lista con cada uno de los salarios.

Además, la división en clases nos sirve para diseñar gráficos importantes, tales como el histograma, las barras, la torta y otros similares

Estructurar una Distribución de Frecuencias

El cuadro que aparece en la pantalla de resultados nos muestra el número total de casos, 474, y las clases salariales que habíamos estructurado.

La clasificación de los salarios en clases, con los rangos determinados, nos permite saber cuántas personas hay en cada clase y graficar con mayor claridad los datos.

Ahora queremos diseñar el cuadro de distribución de probabilidad de frecuencias

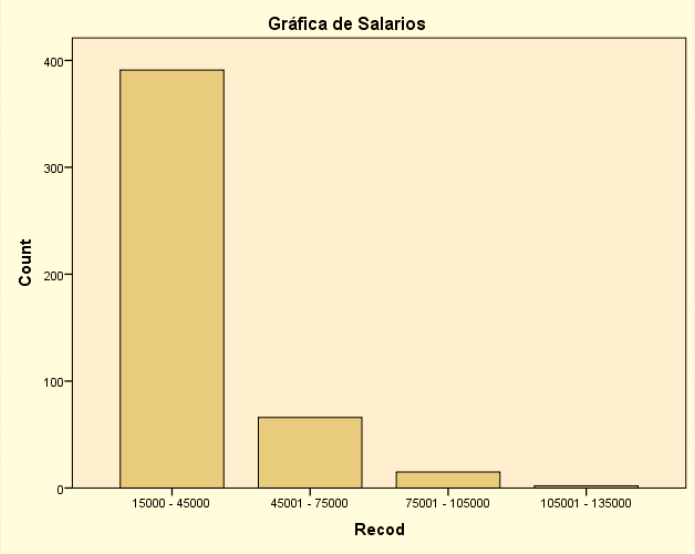
Esto se logra anulando la última columna de la tabla 3.1, haciendo click con el botón derecho del mouse en el encabezamiento que dice "% acumulado".

Pero es necesario visualizar los datos sistematizados en clases.
Eso es posible mediante la opción Gráficas

Para llevar un informe más claro a los ejecutivos, recurrimos a la gráfica de barras.
Para obtener las barras de frecuencias, pulsamos *Gráficas* en el Menú Principal.

Legacy Dialogs → Bar → Simple → Define → Recod (rsalay) → Titles → Gráfica de salarios actuales → Continue → OK

Gráfica 3.2



A diferencia del Histograma, la Gráfica de Salarios registra, explícitamente, las clases que se han estructurado para los grupos salariales

Gráfico de Sectores (Pie)

Usando el mismo procedimiento logramos la gráfica de Sectores o Pie o cualquier otro que figura en la pantalla de gráficas.

→ Graphs → Legacy Dialogs → Pie → Define

Cuadro 3.8

Con la flecha de dirección, arrastramos Recode(rsaly) a la casilla Define Slices by:
Luego → Tiles → Gráfica de Salarios Actales

Gráfica 3.3

Con esto concluimos la primera parte del uso de gráficas; más adelante recurriremos a éstas y otras con diferentes significados y grados complementarios de utilidad

Hagamos una breve pausa para observar cómo vamos aprendiendo Estadística y, al mismo tiempo, el uso del SPSS.

Me pareció más conveniente este método de "aprender sobre la marcha" en vez de dar un curso completo de cada dimensión, una a la vez, por separado.

Cuando terminemos el curso de Estadístico, en sus tres niveles, éste es el primero de ellos, habremos aprendido mucho sobre la utilidad y el manejo del SPSS.

4
PROBABILIDAD

Introducción

La probabilidad es la frecuencia de un suceso determinado que se logra por un experimento aleatorio del que se conocen los resultados posibles, bajo condiciones *dadas.*

Es la posibilidad cuantificada de que algo suceda.

Aunque para resolver todos los problemas que se nos presenten acudiremos al SPSS, sin embargo, es necesario familiarizarnos con la clase de problemas a solucionar.

Los precursores del cálculo de probabilidades fueron Jacob Bernoulli (1674-1705) Thomas Bayes (1702-1761) Joseph Lagrange (1736-1813) y Carl Friedrich Gauss.

La teoría de la probabilidad es la base de las investigaciones estadísticas en las investigaciones de las ciencias sociales y en la toma de decisiones.

En realidad, las llamadas "leyes" en las ciencias sociales son tendencias estadísticas avaladas por una teoría, las que pueden ser estimados con un grado de probabilidad.

Conceptos básicos

Evento: Uno de los posibles resultados de hacer algo.

Si lanzamos una moneda al aire, saldrá "cruz" o "cara".
Cada resultado será un evento

Experimento, la actividad que produce un evento; en este caso, el lanzar la moneda.
¿Cuál será la probabilidad de que una moneda, al ser lanzada, caiga "cara"? será 0,5

Espacio muestral: al lanzar la moneda el espacio muestral es: {cara, cruz}
Si dos eventos pueden ocurrir al mismo tiempo, serán *no mutuamente excluyentes*
Si no pueden ocurrir simultáneamente, serán *mutuamente excluyentes.*

La probabilidad de sacar una carta de un paquete de 52 cartas, será 1/52
La de sacar una reina será 4/52, pues existen cuatro reinas en el mazo.

La de sacar un trébol será 13/52, pues hay 13 tréboles en un mazo.
La probabilidad de sacar una carta roja es 26/52, dado que hay 26 cartas rojas.

La probabilidad de sacar un "as" al lanzar un dado es 1/6, porque hay seis números y un solo "As": del mismo modo con los otros números.

Probabilidad Clásica

La probabilidad de que un evento ocurra es definida del siguiente modo:

$$E = F/(T)$$

E = Evento
F = número de casos favorables
T = el total de casos

En las cartas, el caso favorable de sacar una reina es 4 y el total de casos es 52.
En los dados, el número favorable de sacar un "cuatro" es 1 y el total de casos es 6.

La probabilidad clásica es conocida también como probabilidad a priori.

Se denominaría así, porque las probabilidades de los resultados puede ser conocidos de antemano, tal como sucede con los experimentos de las cartas o los dados.

Frecuencia relativa de la presentación

Es el porcentaje del resultado de casos favorables con relación al total de casos.

Probabilidades subjetivas

Se basan en las creencias de las personas que diseñan el experimento
Es útil cuando no hay antecedentes para una probabilidad objetiva.

¿Cuál es la probabilidad de que el colisionador de partículas fracase? dado que no existe un antecedente, se recurrirá a las suposiciones y al sentido común.

Los responsables de tomar decisiones en una empresa usan la subjetividad para los casos únicos que se presentan a diario en asuntos de precios, y otros similares.

Reglas de la Probabilidad

Los siguientes símbolos son los que se utilizan en el cálculo de probabilidades:

P(A) = Es la probabilidad de que el evento A suceda.
Si puede llevarse a cabo sólo un evento, la probabilidad será sencilla.

Este tipo de probabilidad es conocido como *probabilidad marginal o incondicional*.

Si hay un sorteo para ganar un premio y el total de casos es 60, la probabilidad de que alguien saque el número premiado es 1/60 = 0,0167, sólo un participante podrá ganar.

Eventos mutuamente excluyentes
Hay casos en los que pueden realizarse dos eventos: uno o el otro; supongamos que se tien 5 candidatos para un cargo público y que todos tienen los mismos méritos.

Utilizando el concepto marginal de probabilidad, diremos que la probabilidad de que uno de ellos sea elegido será 1/5.

Pero si estamos interesados en saber la probabilidad de dos candidatos, tendremos que obrar de una manera distinta, en cuanto a la forma, no al contenido.

Supongamos que deseamos saber las probabilidades de que Juan o María ganen el concurso de méritos; es decir, la probabilidad de que alguno de los dos gane.

En este caso tenemos dos eventos que se suman entre sí.

La probabilidad de Juan es de 1/5 y la de María también es 1/5; entonces la probabilidad de que alguno de los dos sea elegido será 1/5 + 1/5 = 2/5 = 0,40

En el caso de que uno de los cinco gane, la probabilidad será: P(A) = 1/5 = 0,20

La probabilidad de Juan o María se representará del siguiente modo: P(A o B) notación que nos indica la probabilidad de que uno de los dos gane el concurso.

Para mostrar gráficamente lo que la suma de probabilidades representa, los teóricos recurren a los símbolos de los conjuntos en matemáticas.

Esos símbolos son muy útiles en la tarea de comprender los teoremas

Tomemos la siguiente tabla, del libro de Levin y Rubin, en la que se consignan datos sobre el número de hijos y sus probabilidades respectivas en una encuesta familiar.

Tabla 4.1

Número de hijos	0	1	2	3	4	5	6
Probabilidad de familias que tienen esa cantidad	0.05	0.10	0.30	0.25	0.15	0.10	0.05

En la tabla 4.1 tenemos una muestra que nos permitirá establecer las probabilidades de que una familia tenga un número determinado de hijos.

De acuerdo con la tabla 4.1, la probabilidad de que una familia tenga 3 hijos es 0.25.

La probabilidad de que una familia no tenga hijos es 0.05, mientras que la probabilidad de que una familia tenga 2 hijos será 0.30 y así sucesivamente.

Ahora aplicaremos estos conceptos: deseamos saber la probabilidad de que una familia del pueblo donde se hizo la encuesta tenga 4 o más hijos.

Nos damos cuenta de que ya no estamos hablando de un solo evento, sino de varios.

Probabilidad de varios eventos

Para plantear el problema recordamos que estamos hablando de varios eventos y que la simbología para representar esa condición es P(A o B)

En el caso que nos interesa, esa expresión toma la forma numérica siguiente:

$$P(4,5,6) = P(4) + P(5) + P(6 \text{ o más}) = 0.5 + 0.10 + 0.05 = 0.30$$

Interpretamos el resultado: la probabilidad de que una familia tenga 4, 5, 6 o más hijos es la suma de las probabilidades marginales de cada evento, esto es, 0.30

Ahora ingresamos a otras dimensiones

Supongamos que deseamos obtener un "diez" o un "trébol" de un mazo de 52 cartas.

Debemos tener en cuenta que también podemos sacar un "diez de trébol".

Vemos que sacar un "diez" o un "trébol" no son eventos mutuamente excluyentes, debido a que hay la probabilidad conjunta de un diez y un trébol al mismo tiempo.

En este tipo de problemas debemos ajustar la ecuación para evitar el conteo doble.
De este modo tendremos: P(diez) + P(trébol) – P(diez de trébol)

$$4/52 + 13/52 – 1/52 = 16/52 = 4/13$$

Otro ejemplo; los empleados de la empresa han elegido a 5 de ellos para que los representen en el consejo de administración; los perfiles de los elegidos quedan registrados en la tabla 4.2

Una vez elegidos, los 5 deciden, a su vez, elegir un portavoz: ¿Cuál será la probabilidad de que la persona elegida sea mujer o tenga una edad por encima de 35 años?

P(mujer o mayor de 35) = P(mujer) + P(mayor a 35) – P(mujer y mayor a 35)

La probabilidad P(Mayor a 35 años) se refiere al total de todos, hombres y mujeres, que tienen más de 35 años; hay solamente dos casos de los cinco: 45 y 40.

Tabla 4.2

Género	Edad
Hombre	30
Hombre	32
Mujer	45
Mujer	20
Hombre	40

$$2/5 + 2/5 - 1/5 = 3/5 = 0.60$$

Probabilidad bajo condiciones de independencia estadística

En primer lugar, definiremos el concepto de "Independencia"

Independencia

Dos eventos serán estadísticamente independientes entre sí cuando el evento o resultado de uno de ellos no tenga influencia en el resultado o evento del otro.

Existen tres tipos de probabilidad independiente: Marginal, Conjunta, Condicional

Probabilidades marginales en condiciones de independencia

Vimos que una probabilidad es marginal o incondicional cuando es la representación simple de un evento; vg. el lanzamiento de una moneda normal.

Ese "experimento" tendrá un evento: cara o cruz, con una probabilidad de 0,5 c/u.

No importa cuántas veces lancemos la moneda, la probabilidad de que salga cara o cruz será la misma; cada lanzamiento es único y no tiene influencia sobre el próximo.

Probabilidades conjuntas bajo condiciones de independencia estadística

La probabilidad de dos o más eventos independientes que se presentan juntos es igual al producto de sus probabilidades marginales.

Representamos ese caso de la siguiente manera: P(AB) = P(A) x P(B)
P(AB) = probabilidad de que los eventos se presenten juntos o probabilidad de A y B

P(A) = probabilidad marginal de que se presente el evento A.
P(B) = probabilidad marginal de que se presente el evento B.

Ejemplos

¿Si lanzamos una moneda por tres veces sconsecutivas ¿Cuál será la probabilidad cruz, cara, cruz, en ese orden?

El resultado será: 0.5 x 0.5 x 0.5 = 0.125, debido a que la probabilidad de que salga cara (A) es independiente a la probabilidad de que salga cruz (B)

¿Cuál es la probabilidad de obtener cruz, cruz y cara, en ese orden luego de tres lanzamientos consecutivos? La probabilidad será: 0.5 x 0.5 x 0.5 = 0.125.

¿Cuál es la probabilidad de obtener al menos dos caras en 3 lanzamientos seguidos? Aquí estamos ante el caso mixto de eventos mutuamente excluyentes.

Esas probabilidades son aditivas, es decir, resultan de la probabilidad de obtener una cara más la probabilidad de obtener otra cara

Pero, tenemos un evento independiente.

Para visualizar este proceso, tomemos la tabla 4.2 del texto de Levin y Rubin, en la que se descomponen las probabilidades de los eventos del total del experimento.

Denominaremos cara = H; cruz = T
En la parte superior del cuadro se registra los tres lanzamientos de la moneda.

Probabilidad de lanzamientos
Tabla 4.2

Un lanzamiento		Dos lanzamientos		Tres lanzamientos	
Resultados posibles	Probabilidad	Resultados posibles	Probabilidad	Resultados posibles	Probabilidad
H_1 T_1	0.5 0.5 1.0	$H_1\ H_2$ $H_1 T_2$ $T_1\ H_2$ $T_1\ T_2$	0.25 0.25 0:25 0.25 1.00	$H_1\ H_2 H_3$ $H_1 H_2 T_3$ $H_1\ T_2 H_3$ $H_1 T_2 T_3$	0,125 0,125 0,125 0.125 1.000
				$T_1 H_2 H_3$ $T_1 H_2 T_3$ $T_1\ T_2 H_3$ $T_1 T_2 T_3$	0,125 0.125 0.125 0.125 1.000

En el primer lanzamiento los posibles resultados son o una cara (H_1) o una cruz (T_1)
La probabilidad para cada uno de estos resultados, tal como vimos es 0.5.

1 y 2 representan el primer y el segundo lanzamiento, en todos los casos.

Lanzada la moneda por segunda vez, nos trasladamos a la columna "Dos lanzamientos" los resultados posibles son: el primer lanzamiento (H_1) el segundo también (H_2)

Éste es un ejemplo de lo que hace la computadora con el SPSS.

Tomamos nota que los eventos del segundo lanzamiento están ligados a los eventos que resultaron en el primero, en cada caso; los mismo con el tercer y el segundo...

El segundo evento posible en la columna del segundo lanzamiento es H_1, T_2, esto es: en el primer lanzamiento salió una cara (H_1) y en el segundo, una cruz (T_1).

El tercer evento muestra que en el primer lanzamiento se T_1 y en el segundo, H_2
El cuarto evento muestra que en el primer lanzamiento T_1 y en el segundo, cara H_2

Las probabilidades en cada caso son 0.25, que resultan de

Cara = probabilidad de 0.5
Cruz = probabilidad de 0.5

La probabilidad del primer evento del segundo lanzamiento H_1, H_2 es 0.5 x 0.5 = 0.25
Lo mismo con las demás probabilidades.

Los datos del tercer lanzamiento se registran de la misma forma; en el primer lanzamiento se tuvo una cara (H_1) en el segundo también (H_2) y en el tercero, una cruz (T_3)

La probabilidad de este evento será 0.5 x 0.5 x 0.5 = 0,125

El mismo razonamiento para los siguientes eventos, hasta que agotamos todas las probabilidades posibles de los tres lanzamientos de la moneda.

Ahora ya podemos responder a la pregunta que nos hicimos al iniciar este capítulo:

¿Cuál es la probabilidad de obtener cruz, cruz y cara, en ese orden, luego de tres lanzamientos?

La pregunta ya nos hace saber que se trata de un experimento de tres lanzamientos
En nuestra tabla vemos que los eventos que la pregunta exige son: T_1, T_2, H_3 = 0.125

Estos ejercicios nos muestran lo que hace la computadora, con el SPSS, cuando le pedimos que calcule las probabilidades de un problema determinado.

Probabilidades condicionales bajo independencia estadística

Vimos dos clases de probabilidad: la probabilidad marginal (o incondicional) y la probabilidad conjunta; la primera representada por P(A) y la conjunta por P(AB)

La Probabilidad Condicional que analizaremos ahora se representa por P(B/A) que muestra dos eventos: A, y B.

De esta manera, la Probabilidad Condicional P(B/A) representa el caso en que el segundo evento B ocurre luego que el primero, A, ya ha tenido lugar.

Nos dice cuál será la probabilidad del evento B una vez que el evento A ya ocurrió.

Antes de continuar, recordemos que para dos *eventos independientes*, A y B, la ocurrencia del evento A nada tiene que ver con el la ocurrencia del evento B.

Tabla 4.3: Probabilidades Condicionales

Tipo de Probabilidad	Símbolo	Fórmula bajo independencia estadística	Fórmula bajo dependencia estadística
Marginal	P(A)	P(A)	P(A/B) x (P(B)
Conjunta	P(AB)	P(A) x P(B)	P(B/A) x P(A)
Condicional	P(B/A)	P(B)	$\frac{P(BA)}{P(B)}$
	P(A/B)	P(A)	$\frac{P(AB)}{P(B)}$

La probabilidad de lograr una cara en un segundo lanzamiento, después de lanzado el primero, seguirá siendo 0.5, porque son eventos independientes. A continuación va una ayuda-memoria para eventos estadísticamente independientes
No olvidemos que la probabilidad marginal, llamada también "incondicional" es (PA)

Probabilidad Condicional Bajo Dependencia Estadística
Antes de proponer la definición formal, vayamos a un ejemplo ilustrativo.

Tabla 4.3: La distribución de las diez bolas

Evento	Probabilidad del Evento
1	0.1
2	0.1 (De color y con puntos)
3	0,1
4	0.1 (De color y con franjas)
5	0.1 (Grises y con puntos)
6	0.1
7	0.1
8	0.1 (Grises y con franjas)
9	0.1
10	0.1

Hay una caja que contiene diez bolas de colores, distribuidas del modo siguiente, tal como aparece en la Tabla 4.3

Tres bolas son de color y tienen puntos
Una es de color y tiene franjas
Dos son grises y tienen puntos
Cuatro son grises y tienen franjas

Siguiendo a Levin y Rubin, hacemos un cuadro para visualizar las condiciones del problema; hay 10 bolas, la probabilidad de sacar una de ellas es 1/10 = 0.10.

Supongamos que alguien saca una bola de color
¿Cuál es la probabilidad de que tenga puntos?

Simbólicamente, el problema puede representarse como P(D/C) es decir:
¿Cuál es la probabilidad de que la bola tenga puntos (D) dado que es de color (C)?

(Vemos que D representa una bola con puntos, C, de color)

Queremos saber la probabilidad de que, siendo la bola de color, que también tenga puntos; para ello, ignoramos las bolas grises, pues no cumplen la condición dada.

Sólo tomaremos en cuenta las que restan.
Hay cuatro bolas de color, tres de las cuales tienen puntos.

Las 4 bolas de color son nuestro "Universo"
Las tres bolas con puntos constituyen el evento

Por lo tanto, la probabilidad de una bola a color con puntos es P(D/C) = ¾ = 0.75
Por otro lado, si nos fijamos en la tabla 4.4 hay una sola bola de color y con franjas.

Por lo tanto, la probabilidad de color con franjas es = ¼ = 0.25

Ambas probabilidades suman 1.
En el próximo capítulo veremos la fórmula general de la probabilidad condicional.

Algunos ejemplos adicionales

¿Cuál es la probabilidad de que la bola tenga puntos, dado que es de color?

Esta pregunta equivale: ¿Cuál es la probabilidad de que la bola sea de color y tenga puntos (son 3 casos de un total de 10) entre la probabilidad de que sea de color.

Probabilidad de que la bola sea de color y tenga puntos = 3/10 = 0.3
Probabilidad de que la bola sea de color = 4/10 = 0.4

La probabilidad de que la bola tenga puntos, dado que es de color, es = 0.3/0.4
Continuamos los datos de la tabla 4.4

¿Cuál es la probabilidad de que la bola tenga puntos, dado que es gris?
Repasando el contenido de la tabla, vemos que:

La probabilidad de bolas grises con puntos = 2/10 = 1/5 = 0.20
La probabilidad de que sea gris es 6/10 = 3/5 = 0.60

La probabilidad de que la bola tenga puntos dado que es gris es = 0.20/0.60 = 0.33

Probabilidades marginales bajo dependencia estadística

Estas probabilidades se calculan sumando las probabilidades de todos los eventos del experimento; v.g: calcular la probabilidad marginal del evento bola de color.

Sumamos la probabilidad de los eventos en los que aparecen bolas de color.

Volviendo a nuestra tabla, vemos que las bolas de color aparecen "bolas de color con puntos" y "bolas de color con franjas"; sumamos las dos probabilidades

$$3/10 + 1/10 = 4/10 = 0.4$$

La probabilidad de una bola gris resultará de la suma de "bolas grises con puntos" y "bolas grises con franjas" es:

$$2/10 + 4/10 = 6/10 = 0.6$$

Distribución de probabilidades

Antes de ir al SPSS haremos una gráfica manual sobre los datos de la tabla 4.5

Construiremos la tabla que sugiere Levin y Rubin para registrar los posibles eventos que resultarían del lanzamiento de una moneda dos veces consecutivas.

"H" significa "cara" y "T" significa "cruz" y las probabilidades son teóricas, puesto que en la realidad puede que en dos lanzamientos los resultados sean diferentes.

Tabla 4.4

Número de cruces	Lanzamiento	Probabilidad del resultado P(T)
	(H, H)	0.25
	(T, H)	0.25
	(H, T)	0.25
	(T, T)	0.25

Pero, si lanzamos una moneda no alterada muchas veces, los resultados obtenidos se irán acercando cada vez más a las probabilidades teóricas.

Deduciremos del anterior cuadro otro que registre la distribución de la probabilidad del número posible de cruces que se obtiene al lanzar dos veces una moneda.

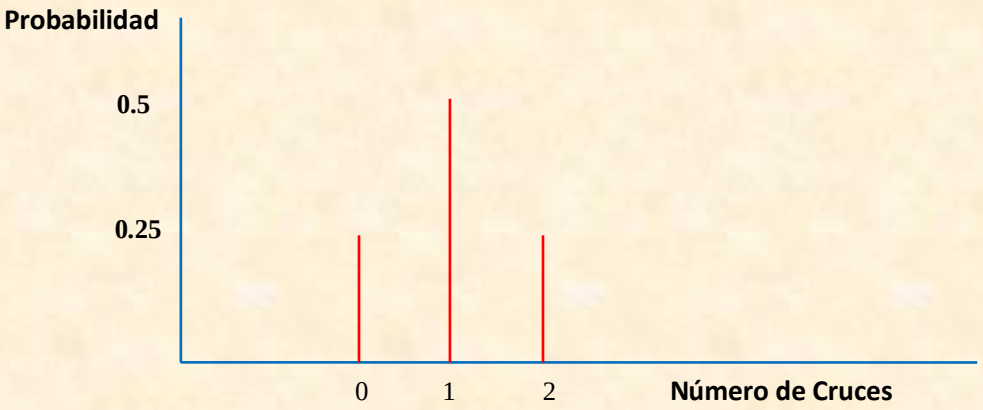
Representaremos gráficamente los resultados de la tabla 4.4, en la gráfica 4.1; estamos interesados en mostrar la distribución de los resultados referidos a las cruces.

Con ese objetivo, colocamos en el eje de las abscisas de un cuadrante el número teórico de cruces que observaríamos en dos lanzamientos de la moneda.

En las ordenadas registramos la probabilidad de cada observación, de acuerdo con los datos de la tabla 4.6 recordemos que estas probabilidades, son teóricas.

En la *gráfica 4.1* reflejamos los datos que nos da la tabla 4.6; ambas, la tabla y la gráfica son dos formas de mostrar una distribución de probabilidades. No olvidemos que la probabilidad de que salga una cruz es (T, H) + (H, T) = 0.50

Gráfica 4.1



Variable Aleatoria

Es la variable que toma diferentes valores como resultado de un experimento aleatorio; puede ser discreta o continua, de acuerdo con el número de valores que toma.

Este ejercicio es útil para tener una imagen de lo que hace el SPSS en la computadora. Definamos algunos conceptos, antes de resolverlo.

En la *Tabla 4.6* registramos los datos de los pacientes atendidos diariamente, durante cien días en una clínica; vamos a convertir esos datos en una tabla de probabilidades

Si toma sólo un número limitado de valores, será una Variable Aleatoria Discreta
Si toma cualquier valor dentro de un intervalo, será una Variable Aleatoria Continua.

En la tabla 4.6 anotamos el número de pacientes mujeres atendidas diariamente en una clínica en los últimos cien días; esto es, una distribución de frecuencias.

Ahora convertimos los datos de las frecuencias en probabilidades
Esos valores son registrados en la *Tabla 4.7.*

Tabla 4.5
Frecuencia de pacientes atendidos en un hospital (Días)

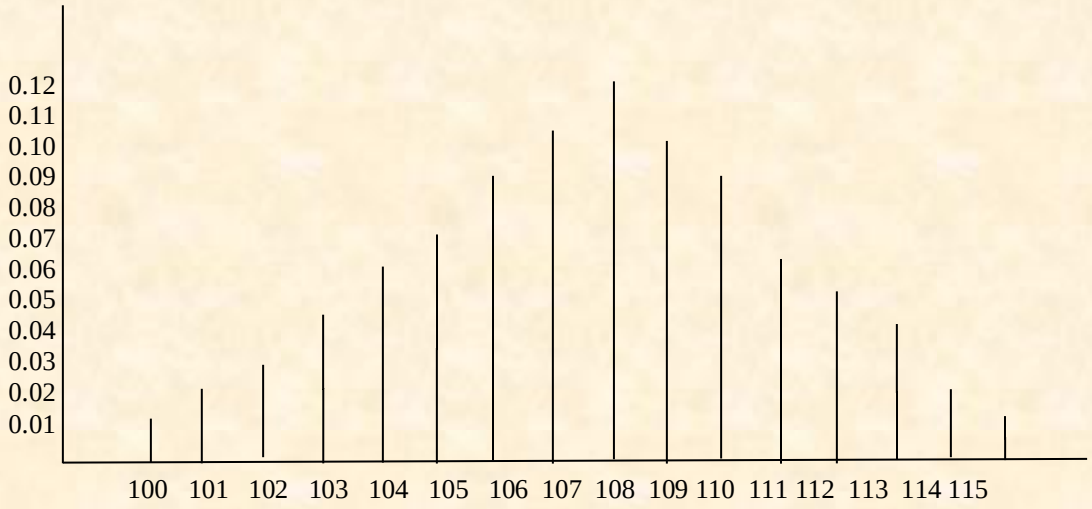
Número de Pacientes atendidos	Número de días que se Atendió
100	1
101	2
102	3
103	5
104	6
105	7
106	9
107	10
108	12
109	11
110	9
111	8
112	6
113	5
114	4
115	2
	100

Tabla 4.6
Distribución de probabilidades del número pacientes atendidos

Pacientes atendidos	Probabilidad de que tome ese valor
100	0.01
101	0.02
102	0.03
103	0.05
104	0.06
105	0.07
106	0.09
107	0.10
108	0.12
109	0.11
110	0.09
111	0.08
112	0.06
113	0.05
114	0.04
115	0.02
	1.00

En la columna izquierda registramos el número de pacientes que son atendidos diariamente en la clínica; en la derecha, se anotan las probabilidades para cada caso.

Gráfica 4.2
Distribución de Frecuencias de probabilidades



Para deducir las probabilidades, se ha dividido el número de días para cada número de pacientes de la tabla 4.7 entre cien, que es el número total de días registrados.

La suma de las frecuencias del número de días anotados en la tabla 4.6 suma 100.
Los datos convertidos en probabilidades en la tabla 4.7, suman en total 1.

La suma de todas las probabilidades siempre es 1, en cualquier caso.
Con los datos de la tabla 4.7 diseñamos una gráfica de *distribución de probabilidad*.

El gráfico 4.2 es para una distribución de probabilidades para variables discretas.

En esa gráfica registramos en el eje de las abscisas el número diario de pacientes atendidas y en el eje de las ordenadas sus respectivas probabilidades.

No dejamos de notar que la longitud más larga corresponde a la probabilidad mayor de la distribución; también notamos la simetría de las longitudes a ambos lados.

Sobre esa simetría es que Gauss llegó a establecer la Distribución Normal para valores continuos, que es la que más a menudo usaremos en el SPSS, una definida.

Hacia la Curva Normal

La Curva Normal, lo dijimos ya, refleja la distribución normal de los datos continuos y es la que los estadísticos tratan de encontrar para realizar sus análisis.

Cuando los datos originales no muestran una distribución normal, los estadísticos transforman los valores de la muestra por otros.

La distribución normal fue presentada por primera vez por Abraham de Moivre en un artículo escrito en 1773

El nombre de "Distribución Normal" fue dado por Charles S. Peirce, Francis Galton y Wilhelm Lexis en 1875

Gauss demostró rigurosamente la distribución normal de los errores y su nombre ha sido asociado a esta distribución, con el nombre de "La Campana de Gauss"

La Curva Normal, o "La Campana de Gauss", refleja la distribución normal de los datos y es la que los estadísticos tratan de encontrar para realizar sus análisis.

Cuando los datos originales no tienen una distribución normal, los estadísticos transforman los valores por otros; vg, en los logaritmos de las variables originales

Valor Esperado de una variable Aleatoria Discreta

Se calcula multiplicando cada valor que la variable pueda tomar, por la probabilidad de que ese valor se presente; al final, sumamos total de estos resultados parciales.

Construiremos un cuadro del *Valor Esperado* de la variable discreta "Número de Pacientes"; para ello, tomamos los valores de la *tabla 4.7* y estructuramos la tabla *4.8.*

El Valor Esperado de la Variable aleatoria "número de pacientes atendidos diariamente durante una muestra de cien días" es 108.20.

Este resultado quiere decir que la clínica, en circunstancias normales, esperará atender un promedio de 108.20 pacientes por día.

Tabla 4.7

Número Atendido	Probabilidad de que la variable aleatoria tome ese valor	Valor Esperado
(1)	(2)	(1) x (2)
100	0.01	1.00
101	0.02	2.02
102	0.03	3.06
103	0.05	5.15
104	0.06	6.24
105	0.07	7.35
106	0.09	9.54
107	0.10	10.70
108	0.12	12.96
109	0.11	11.99
110	0.09	9.90
111	0.08	8.88
112	0.06	6.72
113	0.05	5.65
114	0.04	4.56
115	0.02	2.30
		108.20

Otro ejemplo

En la *tabla 4.8* anotamos la probabilidad para cada venta diaria de fruta.

Tabla 4.8: Ventas durante 100 días

Ventas diarias	Números de días de venta	Probabilidad de venta de cada cantidad
10	15	0.15
11	20	0.20
12	40	0.40
13	25	0.25
	100	1.00

La probabilidad de cada venta diaria se obtiene dividiendo el número de días que se vendió esa cantidad entre 100, que es el total de días que se observó las ventas.

Definición de las clases de pérdidas

Generalmente, las empresas que venden fruta sufren dos clases de pérdidas:

Pérdidas de obsolescencia

Por tener demasiada fruta en un día y botarla al día siguiente.

Pérdidas de oportunidad

*P*or no tener la suficiente fruta para atender a los clientes*;* estas pérdidas se evitan cuando se tiene la cantidad precisa para atender la demanda.

La tabla 4.10 muestra las pérdidas condicionales de nuestro frutero, tanto las pérdidas por la fruta no vendida en el día, como las pérdidas de oportunidad.

Se supone que el frutero ofrece 10 cajas cada día.

Cada caja de fruta le cuesta al frutero $ 20; el frutero vende cada caja a $ 50.
Cuando no hay pérdidas, el cuadro las registra con un 0.

Esos datos son anotados en la tabla 4.10

Tabla 4.10

Posibles demanda	Pérdida Condicional	Probabilidad de demanda	Pérdida Esperada
10	0	0.15	0.00
11	30	0.20	6.00
12	60	0.40	24.00
13	90	0.25	22.50
		1.00	52.50

Todos los valores distintos de 0 representan las pérdidas de oportunidad por los pedidos no cumplidos; si en existencia hay 10 cajas y solicitan 11, perderá $30.

La pérdida de $30 es porque se dejó de ganar $50, a los que se resta $20 que le costó Tal es el caso de la casilla de la segunda fila y la primera columna del cuadro.

Si la demanda es de 13 cajas, pero sólo tiene 10, la pérdida por oportunidad será 90.

La pérdida esperada resulta de multiplicar la "pérdida condicional", por la "Probabilidad de la demanda" para ese día; en la primera fila vemos que no hubo pérdida

¿Qué pasa si el frutero tiene una disponibilidad de 10 cajones pero la demanda es de 13? La pérdida condicional será de $ 90

Si multiplicamos esos $ 90 por la probabilidad de que la demanda sea de 13 cajones, se obtendrá $ 22.50 de pérdida esperada; esto es: 90 x 0.25 = 22.50.

5

DISTRIBUCIÓN DE PROBABILIDADES

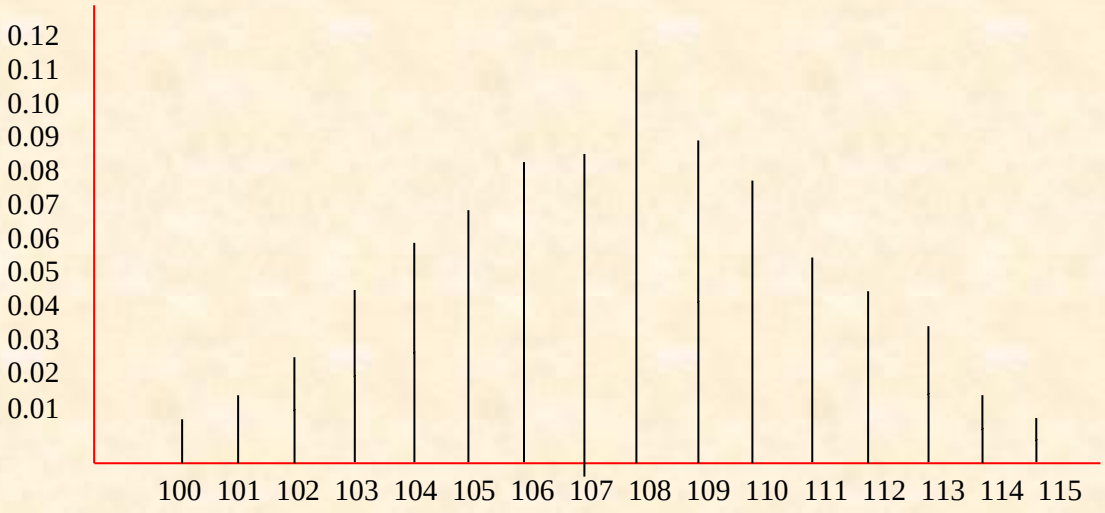
La Distribución Normal

La Distribución Normal es la más importante de todas y la más usada en ciencias sociales; es una distribución continua, no discreta.

Fu diseñada por Karl Gauss en el siglo XIX.
Por eso se llama también distribución o Campana de Gauss.

La Distribución Normal tiene características que son muy útiles cuando es preciso decidir tomando como base las muestras que se levantan de una población dada.

Gráfica 5.1
Distribución de Probabilidad Normal



Representa muy bien las distribuciones de fenómenos reales, incluyendo características humanas, tales como el peso, la altura, el coeficiente de inteligencia y otras.

Con el objeto de mostrar el proceso de cómo se llega a la Curva de la Distribución Normal, reproduzcamos en la gráfica 5.1, la *gráfica 4.2* del capítulo anterior.

La Gráfica 5.1 es una distribución normal para valores discretos.

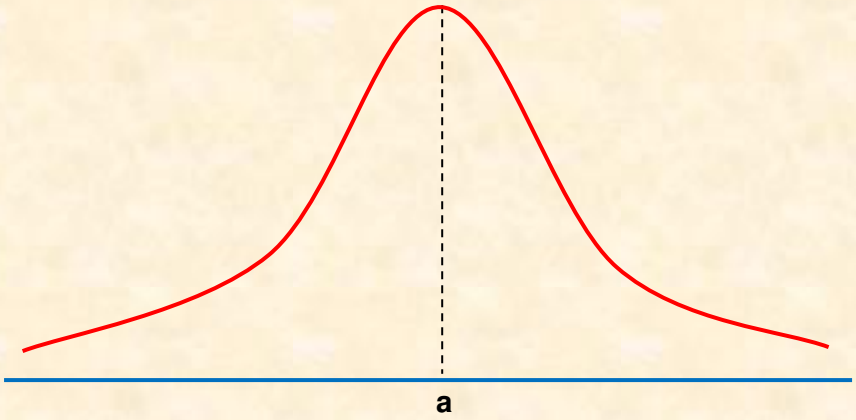
Hagamos que los intervalos entre los valores de las probabilidades registrados en las ordenadas y los intervalos del número de pacientes, se vuelvan infinitesimales.

Es decir, que los valores de vuelvan continuos.
La curva resultante será la *Curva Normal que se muestra en la gráfica 5.2*

La superficie de la Curva Normal está conformada por la adición de todas las probabilidades de los resultados de un experimento.

Precisamente, la tarea más importante de operar con la Curva Normal es encontrar la probabilidad de un resultando como efecto de un experimento.

Gráfica 5.2
Una Curva Normal



Propiedades

Lo dijismo ya, por lo general, el estadístico trabaja con muestras
La Curva normal es muy importante en este aspecto

La Curva Normal tiene un solo pico; es decir, sólo tiene una moda, es unimodal; la Media la Mediana y la Moda, están en el mismo punto central, son iguales entre sí.

Adoptemos la simbología de los parámetros de la Curva Normal:

σ representa la desviación estándar de la curva normal
X* representará la Media de la muestra; también se usa la X con una barra encima.

Los extremos, izquierdo y derecho, se extienden y nunca tocan la línea base.
La mayor densidad de frecuencias está en **a**

En el punto **a**, la Media, la Mediana y la Moda son iguales.

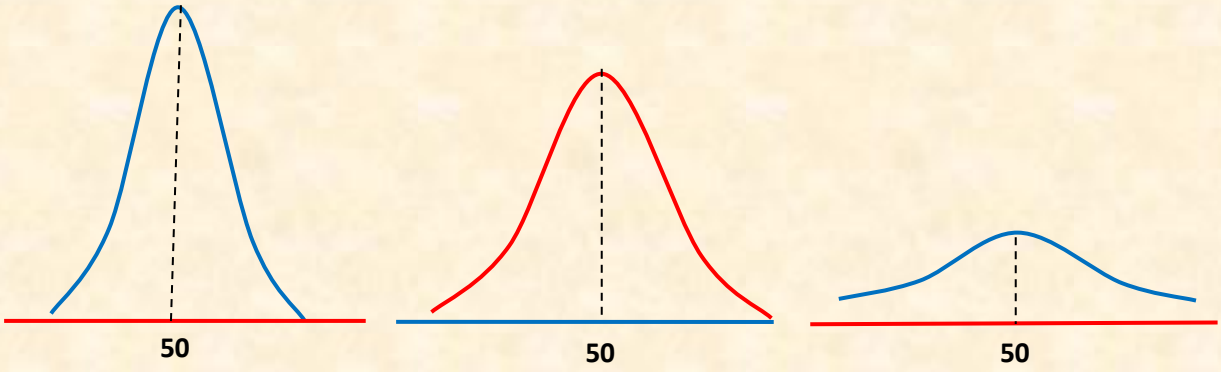
Teórica y prácticamente, no hay una sola curva normal, sino una familia completa. Para definir una curva normal, lo único que necesitamos es definir dos parámetros:

La media = **μ**
Error estándar = **σ**

Conociendo los dos parámetros, se puede diseñar una curva de distribución normal. La gráfica 5.2 muestra curvas normales con diferentes medias.

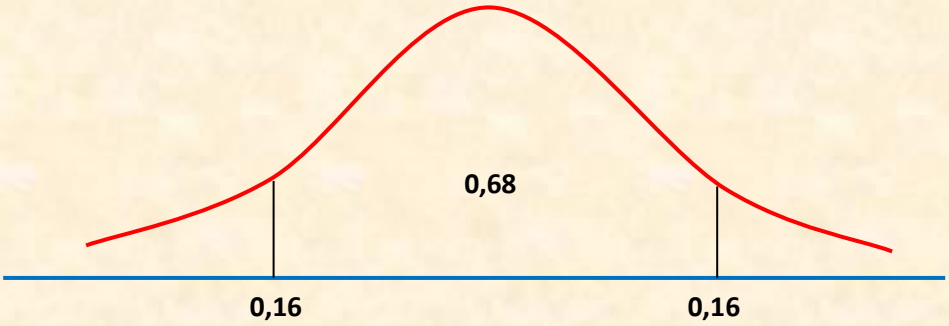
En la siguiente serie de tres curvas normales, la de la izquierda tiene la desviación estándar más pequeña; la del medio, un poco mayor y la tercera tiene la mayor

Gráfica 5.3
Curvas normales



Sin embargo, todas ellas tienen una Media = Mediana = Moda = 50.
La Media, la Mediana y la Moda dividen la curva normal en dos partes iguales.

Gráfica 5.4: Área bajo la curva normal



Área bajo la Curva Normal

El área de cualquier curva normal es 1.00, independientemente de su media o desviación estándar y la superficie es un conjunto completo de probabilidades.

Conociendo que La Curva Normal tiene una media y una desviación estándar, es posible derivan las siguientes características que las singularizan en cada caso.

Alrededor del 68% de los valores de una población normalmente distribuida está dentro de una distancia que equivale a una desviación estándar de la media.

En la *gráfica 5.4* vemos que el 68% de los datos están en el centro; el resto, 32%, se dividen en las dos alas de la izquierda y la derecha, cada una con el 16%.

Por ello, la distancia entre las verticales de esa curva es dos **σ** a la izquierda y dos **σ** a la derecha de las observaciones cuya distribución es normal.

En la *gráfica 5.5*, el 95% de las observaciones están en el área central y el 5% restante se reparten en las alas de la izquierda y derecha, cada una con el 2.5%.

Gráfica 5.5
Área bajo la curva normal

Esto quiere decir que el 95.5% de los valores de una población normalmente distribuida se encuentra dentro de dos errores estándar de la media

En este momento no tenemos aún una idea clara de lo que esto significa, pero a medida que avancemos captaremos la lógica de la curva normal y la manera de usarla.

Con el objeto de familiarizarnos con los conceptos, vamos a proceder con el planteamiento y la solución de algunos ejercicios sobre el uso de la Curva Normal.

Ejercicio 1

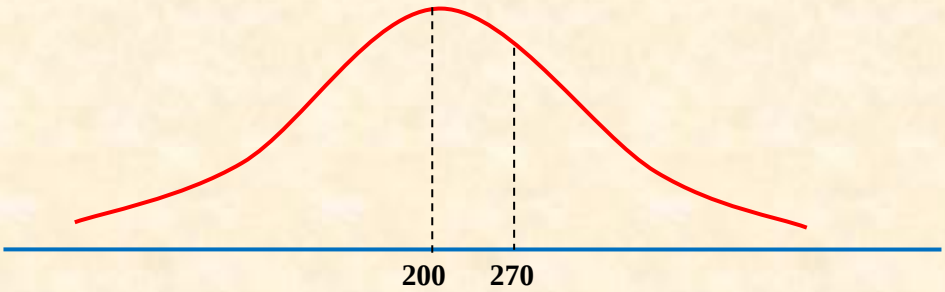
Supongamos que la Media Aritmética (la Media) de los ingresos mensuales de los alumnos que cursan la materia es de $ 200 y que el error estándard es $\sigma = 70$.

Deseamos estimar el monto del ingreso que esté a una distancia de una desviación estándar (70) a la derecha de la media

Con esos datos podemos diseñar la Curva Normal que se muestra en la *gráfica 5.6.* Podemos hacerlo, porque ya sabemos que $\mu = 200$ y $\sigma = 70$.

Según las condiciones del problema, queremos identificar el monto de ingresos que está por encima de la media a una distancia de una desviación estándar.

Gráfica 5.6

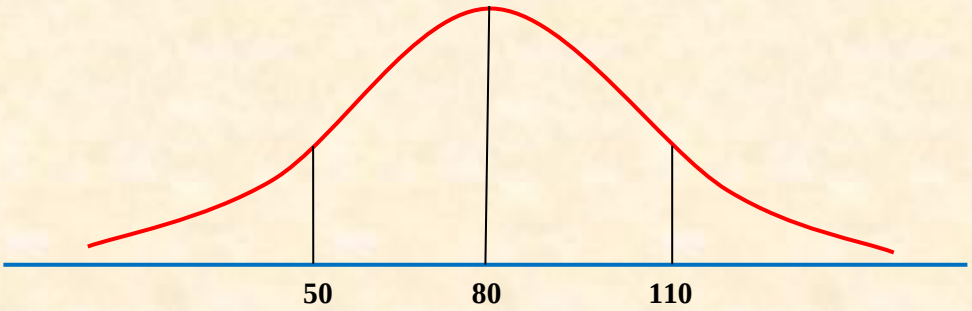


Si la desviación estándar es $\sigma = 70$, entonces la observación a la derecha que está a una distancia de una desviación estándar de la media es 270.

Ejercicio 2

Supongamos que tenemos una curva normal cuya media es $\mu = 80$ y una desviación estándar $\sigma = 10$; nos piden los valores que están a 3 desviaciones estándar de la media.

Gráfica 5.7



La *gráfica 5.7* es la curva normal que tiene la media = 80 y la desviación estándar 10.

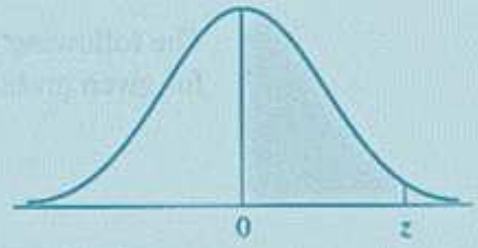
El valor de la derecha es 110: que mide la distancia que la separa de la media en 3 desviaciones estándar. El valor de la izquierda es 50, pues se distancia 3 desviaciones estándar de la media, a la izquierda estas tres desviaciones se restan de la media

Tabla de valores de la Curva Normal Representa el área bajo la curva normal entre la media y cualquier valor que asuma la variable aleatoria normalmente distribuida.

En la primera columna, la tabla registra los valores "z" donde "z" es la desviación estándar de una curva normal; sus valores van desde 0 hasta 3.6 desviaciones estándar.

Las siguientes columnas representan las probabilidades de las respectivas desviaciones estándar, en cada caso.

The following table gives the areas under the standard normal curve from 0 to z.



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999

Ejemplo, el valor 0.2, bajo la columna "z" representa la desviación estándar: $\sigma = 0.2$

Si z = 0.2, la segunda columna muestra el área bajo la curva que es 0.0793; es decir la probabilidad de que algo ocurra, la desviación estándar de Z = 0.2

Las demás columnas muestran los valores decimales de la desviación.

Cuando z tiene más de un decimal, el segundo decimal se representa en la columna en la que se encuentra la probabilidad buscada

Por ejemplo, si la desviación estándar es igual a 0,25, el área respectiva estará bajo la columna "5" y entonces la probabilidad de una desviación de z = 0.25 será 0.0987

No se preocupen, vamos a lograrlo

No olvidemos que la Curva Normal se divide en dos partes exactamente iguales y que el centro está dado por el valor de la Media Aritmética

En el siguiente ejemplo vamos a utilizar el proceso inverso para comprender mejor lo que estamos haciendo con la tabla.

Busquemos un valor cualquiera, digamos 0.3508; ubicado ese valor en la tabla vemos que se encuentra bajo la columna "4" y la fila que corresponde a z = 1.0

En consecuencia, el valor 0.3508 corresponde a: z = 1.4

En palabras más claras: la probabilidad 0.3508 represetna 1.4 desviaciones estándar, es decir, $\boldsymbol{\sigma} = 1.4$ de la media aritmética de una distribución normal

En consecuencia, queda establecido que z es el número de desviaciones estándar que hay entre el valor x que buscamos y la media de la distribución normal.

Los matemáticos han encontrado que el valor z, que es el valor de cada desviación estándar, se define como:

$$z = (x - \mu)/\sigma$$

x = valor de la variable aleatoria que buscamos
$\boldsymbol{\mu}$ = media de la distribución de la variable aleatoria
$\boldsymbol{\sigma}$ = desviación típica de la distribución

Continuemos con los ejemplos manuales
Una vez que tengamos las ideas claras, en otro capítulo acudiremos al SPSS.

Estos ejemplos nos permitirán visualizar el uso de la tabla de áreas bajo la curva normal; recordemos que la curva está dividida en dos partes iguales.

Ejercicio 3

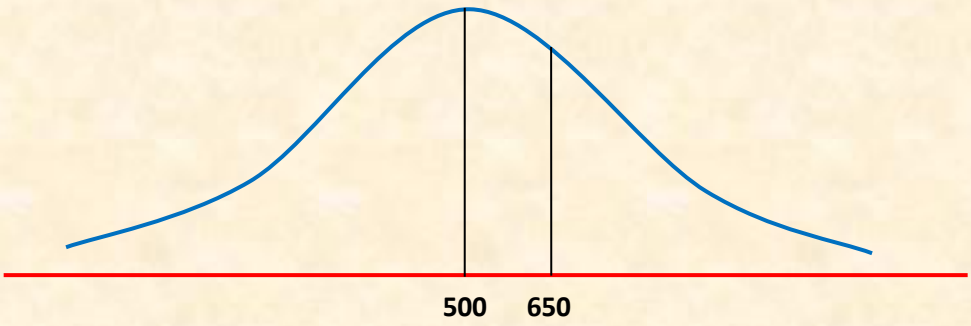
Supongamos que está en proceso un curso de Estadística; al término del curso recogemos los datos sobre el tiempo que tomó a los alumnos completarlo

Además sabemos que la distribución del tiempo que toma a los alumnos aprender el curso, es normal, gráfica 5.8

¿Cuál es la probabilidad de que un participante elegido al azar se tome entre 500 y 650 horas para completarlo, dado que el tiempo promedio de aprendizaje es de 500 horas?

También se conoce que la desviación estándar es de 100 horas
Los términos del problema quedan registrados en nuestra curva normal, Gráfica 5.8

Gráfica 5.8



Lo primero que anotamos son los estadísticos:
La media $\mu = 500$ y la desviación estándar, $\sigma = 100$

Se nos pide la probabilidad de que un alumno tome entre 500 y 650 horas para cubrir el curso; la probabilidad de ese evento se encuentra entre 500 a 650 horas P(500 a 650)

Para resolver el problema debemos estandarizar los datos con nuestra fórmula:

$$z = (x - \mu)/\sigma; \quad z = (650 - 500)/100; \ 1.5$$

Hemos convertido los datos en desviaciones estándar; es decir en, $z = 1.5$.

Esto es, la cantidad de horas cuya probabilidad nos piden = 650 está ubicada a una distancia de $z = 1.5$ desviaciones estándar de la media

Recurrimos a la Tabla de Distribución Normal, la que ya usamos anteriormente, y buscamos en la columna de las $z = 1.5$

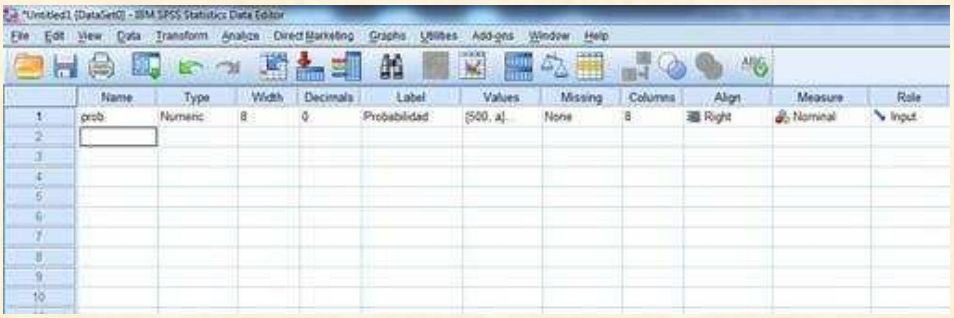
Como no hay más decimales, encontramos el valor que buscamos en la columna inmediata a su derecha (0.00) que es igual a 0.4332

Así, decidimos que la probabilidad de que un participante elegido al azar emplee de 500 a 650 horas para terminar el curso, es 0.4332.

Con el SPSS

Vayamos a la Pantalla Variable View del SPSS
Creamos la variable prob

Pantalla 5.1



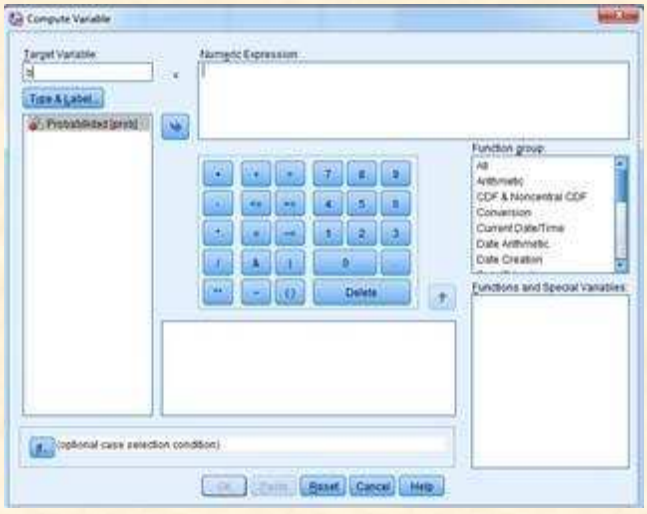
En la columna Valores anotamos

500 =a
650= b

La pantalla 5.1 muestra lo que hemos hecho en la Pantalla Variable View
En la Pantala Data View se ha creado la variable prob, debajo de la cual anotamos 0

Click → Menu → Tranforms → Computer variable → b en la casilla Target Variable
La Caja 5.1 muestra el estado del proceso

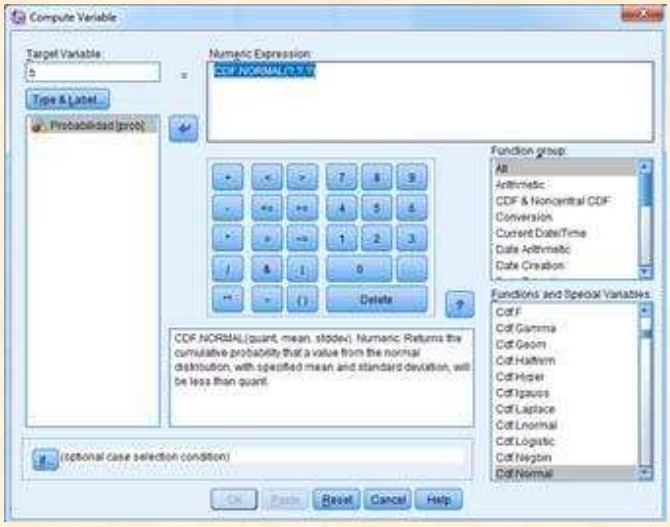
Caja 5.1



En la parte derecha, debajo de la pantalla horizontal hay una pantalla pequeña con el nombre Function group; elegimos *All*

De inmediato en la pantalla inferior se despliega todas las funciones del SPSS
De todas ellas elegimos Cdf.Normal

Caja 5.3



Click a la flecha de dirección arriba; en la pantalla superior aparecer la siguiente expresión: CDF.NORMAL(?,?,?) tal como en la Caja 5.3

La primer interrogante nos pregunta cuál es valor que deseamos estimar
Borramos ese signo y anotamos 650

La segunda interrogación nos pregunta cuál es la media de la distribución sobre la que estamos trabajando, anotamos 500

La tercera interrogación nos pregunta el valor de la Desviación Estándar, ponemos 100
La pantall 5.3 muestra el estado del proceso: CDF.NORMAL(650,500,100)

Anotamos manualmente un signo de resta al final de CDF.NORMAL(650,500,100)-

Retornamos a la pantalla de funciones y traemos nuevamente Cdf.Normal
La subimos con la flecha respectiva

Llenamos con 500, 500, 100 que son los indicadores del problema original
En la pantalla de arriba leemos ahora:

CDF.NORMAL(650,500,100)-CDF.NORMAL(500,500,100) → OK

El SPSS nos lleva a la pantalla Variable View
Allí vemos que se ha creado la variable b

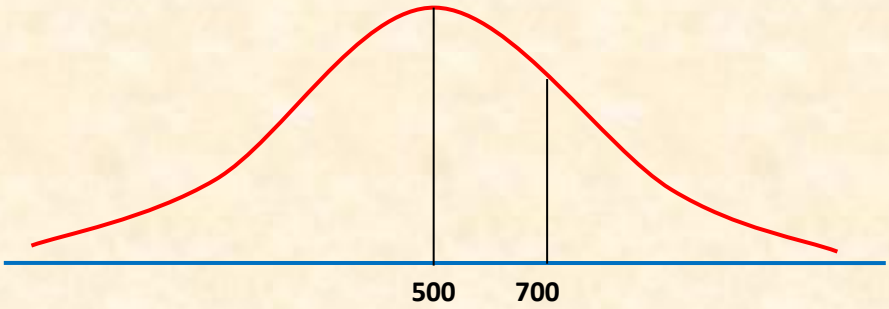
En la columna de los decimales para la nueva variable b, anotamos 6

Vamos a Data View y bajo la columna b encontramos el valor 0.433193
Que es el resultado que habíamos encontrado manualmente

Ejercicio 4

Con los mismos datos del anterior problema ¿Cuál es la probabilidad de que un participante, elegido al azar, se tome más de 700 horas en completar el programa?

Gráfica 5.9



Tomamos nota de los estadísticos dados y diseñamos la gráfica 5.9
x = más de 700 horas; $\mu = 500$; $\sigma = 100$

Diseñamos una curva normal; en ella debemos encontrar la Probabilidad (más de 700)

Lo que nos piden es encontrar la probabilidad a la derecha de 700 horas.
Sabemos que $\mu = 500$; x = más de 700; $\sigma = 100$

Convertimos nuestros datos en desviaciones estándar; $z = (700 - 500)/100 = 2$
700 horas está a 2 desviaciones estándar de la media (la Media es 500)

En la tabla encontramos que para 2 desviaciones estándar, la probabilidad es 0.4772
Pero el problema no se refiere a 700 horas, sino a más de 700 horas

Es decir, no se refiere a la probabilidad que al alumno le tome entre 500 y 700 horas, sino a la probabilidad de que el aprendizaje le tome más de 700 horas.

La gráfica 5.9 nos indica que lo que buscamos es la probabilidad a la derecha de 700
Sabemos que la mitad derecha, como la izquierda, tiene una probabilidad de 0.5

Lo sabemos, porque dijimos que la media aritmética dividía la curva normal en dos partes exactamente iguales, es decir, cada una tiene un valor de 0.5.

Usaremos esa característica.
Ya encontramos que la probabilidad de 700 horas = 0.4772

Ahora realizamos la siguiente operación: valor de la mitad derecha de la curva = 0.5, menos la probabilidad de 700 = 0.4772

Esto es: 0.5000 – 0.4772 = 0.0228

Esto es, la probabilidad de que al participante le lleve más de 700 horas para aprender el curso de Estadística es 0.0228

Con el SPSS

Con las mismas variables, vamos a Vista de Datos y en Valores anotamos

500 = a
700 = b

Con el mismo procedimiento que los anteriores vamos a Transformar, registramos b en Variable de destino y en la pantalla grande anotamos: 1-CDF.NORMAL(700,500,100)

En la pantalla Variable View, se ha creado la variable B, con dos decimales.
Lo aumentamos a 6

El Resultado aparece en la Pantalla Data View: 0.22750, que es el mismo que encontramos manualmente, haciendo uso de la tabla normal

Ejercicio 5

Con los mismos datos del problema original, se desea saber la probabilidad que a un participante le tome entre 550 y 650 horas aprender el curso total, gráfica 5.11.

Nos damos cuenta de que es preciso calcular el valor de dos x: o sea x_1 y x_2
Los datos son $\mu = 500$; $x_1 = 550$; $x_2 = 650$; $\sigma = 100$

Primero calculamos el valor de cada x.

De inmediato notamos que el valor de 550 está a la derecha de la media = 500
Empezamos calculando el número de desviaciones estándar entre 550 y 500

$$z = (x - \mu)/\sigma = (550 - 500)/100 = 0.5$$

Buscamos en la tabla el valor de 0.5 en la columna de las desviaciones estándar (z)
Ese valor es 0.1915

Ahora hacemos lo mismo con el cálculo tomando como x = 650.

$$z= (650 – 500)/100 = 1.5$$

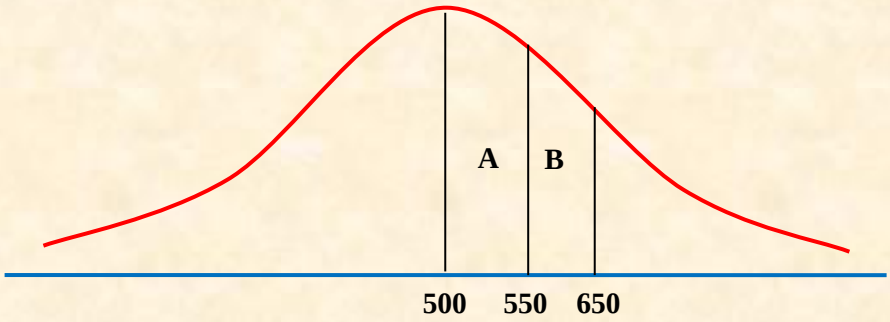
El valor para 1.5 desviaciones estándar en la tabla es 0.4332.
El problema consiste en hallar la probabilidad entre 550 y 650 horas

Esto representa hallar el área entre 550 y 650, tal como se ve en la siguiente curva, recordando que la media es 500.

La gráfica 5.10 muestra la estructura del problema.

Para encontrar el valor que nos pide el problema, que está entre 550 y 650 horas, tomamos el área entre 650 y 500 y le restamos el área entre 550 y 500.

Gráfica 5.10



Es decir, tomamos la distancia total entre 500, la media, y 650 el valor mayor
Esa distancia está representada por la suma de las áreas A + B.

Luego estimamos la distancia entre 550 y 500, que es el área A.
En la tabla de las áreas de la Curva Normal encontramos los siguientes valores:

Área A + B = 0.4332
Área A = 0.1915
Área B = 04332 - 0.1915 = 0.2417

La probabilidad de que un participante necesite un tiempo entre 550 y 650 horas para vencer el curso completo de Estadística, es: 0.4332 – 01915 = 0.2417

Ese valor es el que corresponde al total del área B

Esto es: La probabilidad de que a un participante le tome entre 550 y 650 horas aprender el curso total es 0.2417

Con el SPSS
Vamos a Vista de Variables, creamos la variable *esta, con 6 decimales;* en la casilla de valores asignamos; en la pantalla de datos aparecerá la nueva variable, debajo de la cual anotamos el valor 0

a = 550
b = 650

Después de dar los valores respectivos en la columna Valores, vamos a → Transformar
Anotamos **b** en Variable de Destino

En la pantalla de la derecha registramos:
CDF.NORMAL(650,500,100)-CDF.NORMAL(550,500,100) →O.K

Aparecerá, como en los demás casos una pregunta Change existing variable? → OK

Eso quiere decir que el valor anterior, correspondiente al anterior ejercicio en la pantalla Data View, será cambiado por el nuevo valor estimado

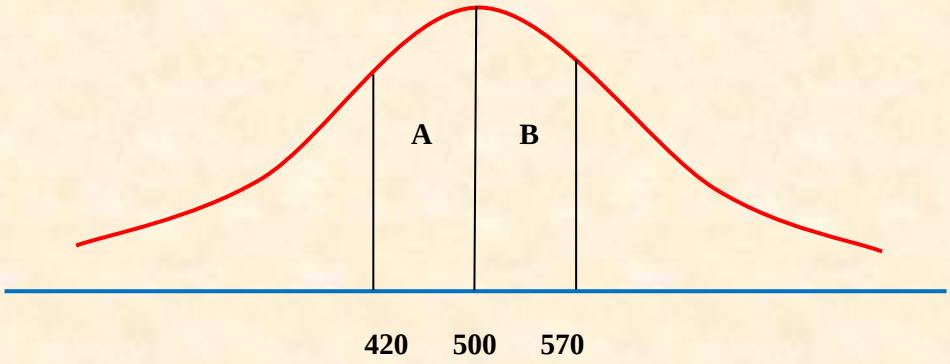
El resultado con 6 decimales será: 0.241730
Que es el mismo que hallamos por el procedimiento manual

Poco a poco, le vamos tomando el pulso al asunto de encontrar las probabilidades.
Por eso seguimos con algunos ejercicios más.

Ejercicio 6
¿Cuál es la probabilidad de que un participante escogido al azar se tome entre 420 y 570 horas para aprobar el curso?

Diseñamos nuestra curva normal con las condiciones requeridas, gráfica 5.12.

Gráfica 5.11



El problema nos pide la suma de las probabilidades que se encuentran entre 420 y 500 más la probabilidad que hay entre 570 y 500.

Necesitamos áreas en las dos mitades de la curva.
Es decir, sumar las áreas A + B

Empezaremos calculando los valores de sus respectivas desviaciones estándar = z.

$z_1 = (420-500)/100 = -0.80$; $z_2 = (570 - 500)/100 = 0.70$

Las desviaciones estándar son -0.80 y 0.70 respectivamente.

No tomamos en cuenta el signo negativo y más bien buscamos las probabilidades respectivas a cada z en la tabla.

Las probabilidades son, respectivamente 0.2881 y 0.2580
Sumamos estas probabilidades: 0.2881 + 0.2580 = 0.5461.

La probabilidad de que un participante escogido al azar requiera entre 420 y 570 horas para aprender el curso de estadística es 0.5461.

Con el SPSS
En Vista de Variables, creamos la variable dif con 6 decimales

Valores: **a** = 420 y **b** = 570
Crear las dos variables y en la pantalla Vista de datos poner **b** debajo de esta

En Transformar, click **b** en Variable de Destino y en la pantalla de la derecha
CDF.NORMAL(570,500,100)-CDF.NORMAL(420,500,100)

En la pantalla Variable view la variable b tiene sólo dos decimales
Aumentamos su número a 6

Resultado = .546181

Ejercicio 7
¿Cuál es la probabilidad de que a un participante le toma más de 500 horas?

Sabemos que la media es 500 horas, en consecuencia, el problema se reduce a encontrar el área de la mitad derecha de la curva normal, esto es, 0.5000

Esa es también la probabilidad de que el participante le tome más de 500 horas.

Lo que hicimos en los anteriores capítulos fue formarnos una idea de cómo se distribuyen las probabilidades de ocurrencia cuando las distribuciones son discretas.

6
OTRAS DISTRIBUCIONES

La Curva Normal es una distribución de variables continuas, es decir, para variables cuyos valores varían infinitesimalmente.

Hay otros modelos teóricos de distribución de probabilidades que los estadísticos usan con gran frecuencia y que no son continuas; esos modelos se encuentran en el SPSS.

Para usarlos con propiedad, debemos captar muy bien el concepto de cada una.

La Distribución Binomial

Es una distribución de variables discretas; v.g, el lanzamiento de una moneda o una distribución en la que haya dos variables: cierto-no cierto; sí-no; éxito o fracaso

Cada lanzamiento de la moneda tiene sólo dos resultados: cara o cruz; sí o no, cada evento tiene un probabilidad de ocurrencia, tal como las que vimos hasta ahora.

Esas probabilidades permanecen fijas en el tiempo
Por añadidura, diremos que los intentos son estadísticamente independientes entre sí.

Las propiedades de una distribución binomial

La Distribución Binomial se caracteriza principalmente, porque resuelve problemas que incluyen respuestas como "sí o no"; "cara o cruz" y otros parecidos.

Para ello toma en cuenta las probabilidades de éxito o de fracaso, las que fueron estimadas por los matemáticos que consolidaron la estadística.

Los teóricos son los que se encargan de demostrar cada fórmula utilizada

Por otro lado, toman en cuenta el número de éxitos deseados y el número de intentos
Así, es que los problemas se plantean generalmente del siguiente modo:

p = probabilidad de tener éxito; 0.5
q = 1 – p, es la probabilidad de fracaso; 0.5

r = número de éxitos deseados.
n = número de intentos hechos.

El tipo de problema que debemos resolver es el que nos indica en que ocasiones recurriremos a la distribución binomial o a otra cualquiera

La Distribución Binomial y el SPSS

Luego de haber captado el concepto respectivo, usemos el SPSS para resolver problemas relacionados con varios tipos de distribución de probabilidades.

Ejercicio 6.1

En una escuela la probabilidad de que un estudiante falte es 0.4
Tomamos una muestra de 5 alumnos

¿Cuáles serán las probabilidades de que, 1, 2, estudiantes falten?

Caso 1: un alumno

Resolveremos el problema caso por caso
Calcularemos la probabilidad de que un estudiante, exactamente, llegue tarde

Usaremos la opción PDF.BINOM

Caja 6.1

En la pantalla Variable view creamos la variable *falta*
En la columna de Valores: damos el valor de 1= a

Transform → Compute variables → **a** *en target value*

En la pantalla de Function group anotamos All
En el menú desplegado con las funciones del SPSS escogemos Pdf.Binom

Click en la flecha de dirección hacia arriba
En la pantalla superior aparecerá PDF.BINOM(?,?,?)

Rellenamos los signos de interrogación con: PDF.BINOM(1,5,0.4) → OK
Cada valor significa:

1 = un estudiante
5 = número de estudiantes tomados en cuenta para el experimeno
0.4 = probabilidad, ya conocida, de que un alumno falte

En la pantalla Variable View aparece la variable **a** con dos decimales
Aumentamos su número a 6

En pantalla de Vista de Datos, bajo la columna a, aparece el número 0.259200
Por lo tanto, la probabilidad de que, exactamente, un estudiante falte es de 0.259200

Solución para 2
b = 2 (La probabilidad de que falte exactamente 2 estudiantes)
Procedemos de la misma manera que en el anterior ejercicio

Pero esta vez, no calculamos la probabilidad de que un alumno falte, sino que calculamos la probabilidad de que falten 2 alumos

En la pantalla de la derecha anotaremos PDF.BINOM(2,5,0.4) → OK
En la Pantalla de Vista de datos tendremos la probabilidad calculada: 0.345600

Esto es, con los datos que se nos proporcionó, poemos establecer que la probabilidad de que 2 alumnos, exactamente, falten es 0.345600

Otras Probabilidades

Supongamos que nos piden estimar la probabilidad de que a lo sumo 2 alumnos falten

Observamos que en el anterior ejercicio se nos pedía que estimáramos la probabilidad de que 2 alumnos, exactamente, faltaran

Ahora se nos pide la estimación que, a lo sumo, 2 alumnos falten
El problema difiere del anterior

Para esto usamos la función Cdf.Binom
Creamos la variable sumo y en la casilla "Valores" registramos: 2 = a

Anotamos a en la casilla Target Variable
En el menú de funciones, pulsamos Cdf.Binom = CDF.BINOM(2,5,0.4) → OK

Enla Pantalla de Vista de datos tendremos la probabilidad calculada: 0.68 para dos decimales; aumentamos el número de decimales de a hasta 6, el resultado será: 0.682560

La probabilidad de que, a lo sumo, 2 alumnos falten es 0.682560

Ejercicio 6.2

Se nos pide estimar la probabilidad de que en tres lanzamientos de una moneda aparezcan dos caras

Solución

En la pantalla Variable View creamos la variable event, tal como ya sabemos hacerlo

Dado que en este caso, también nos piden un valor exacto, 2 caras, usaremos la versión *Pdf.Binom,* conociendo que la probabilidad de que salga una cara es 0.5

En la pantalla de variables creamos la variable cara y en la columna "valores" 0.5 = a

Transform → a en taget Variable →PDF.BINOM(2,3,0.5)

Que corresponden a 2 caras, 3 lanzamientos y la probabilidad de que salga cara

El resultado es 0,375000, es decir, la probabilidad de que luego de tres lanzamientos de una moneda, tengamos dos caras en cualquier orden es 0.375000

Ejercicio 6.3

Se quiere calcular la probabilidad de hallar tres tubos llenos de una sustancia química de una muestra de 6 sabiendo que la probabilidad de lograr un tubo lleno es 0.8

Solución

Vamos a la pantalla Variable View y creamos la variable tub sin decimales
En la columna de valores hacemos a = 3

3 = tubos cuya probabilidad de encontrar llenos se desea estimar
6 = el número total de tubos que se tomará en cuenta
0.8 = la probabilidad de encontrar 1 tubo lleno

Transform → target value; a; → PDF.BINOM(3,6,0.8)
En la pantalla Variable View aparece a con 2 decimales, los aumentamos a 6

En la pantalla Data View, en la columna se registra el valor: 0.081920

Esto es: la probabilidad de que tomando una muestra de 6 tubos de ensayo encontremos 3 llenos, sabiendo que la probabilidad de encontrar 1 lleno es 0.8 será: 0.081920

La Distribución de Poisson

Es otra de las distribuciones más usadas para resolver problemas con variables discretas; esta distribución debe su nombre a su creador, Denis Poisson (1781-1840)

Hay varios tipos de problemas que requieren de esta distribución.

Por ejemplo, la distribución de llamadas telefónicas que llegan a una central y las llamadas que los pacientes internados en las clínicas hacen a las enfermeras.

También debemos incluir el número de coches motorizados que llegan a un control de rodaje y otras; todos se refieren a procesos discretos con una o más ocurrencias.

Así, el número cotidiano de llamadas a las enfermeras puede representarse por 1, 2, 3, 4,… lo mismo el de llegadas de vehículos a pagar rodaje.

La diferencia entre distribución de Poisson y la Binomial estriba en lo siguiente:

La Distribución Binomial toma en cuenta dos eventos: un **sí** y un **no**
La Distribución de Poisson incluye más opciones.

Cálculo con el SPSS

El proceso es igual al que empleamos en el cálculo de la distribución binomial.

Ejercicio 6.4

Se desea averiguar cuál es la probabilidad de que suceda 0 accidentes, 1 accidente y 2 accidentes en un día, en una calle donde el promedio es de 5 accidentes por día.

Planteo del problema

En la pantalla de *Vista de variables* creamos una variable, *even.*
En Valores asignamos: 0 = a

En la pantalla de datos aparece la variable even
Colocamos un 0 en la primera casilla

Menú Principal → *Transformar* → *Calcular variable*: allí ingresamos la variable a con el objeto de estimar la probabilidad de que ese día haya cero número de accidentes.

En la parte inferior derecha hay un cuadro de funciones con un menú; pulsamos *Todo*
Del menú de funciones buscamos la función Pdf-Poisson.

Las características de la función Poisson aparecerán debajo del teclado; la subimos y en la pantalla aparece Pdf.Poison (?,?)

Reemplazamos el primer interrogante con 0 y el segundo con 5 → OK
En la pantalla *Vista de datos* aparece el número 0,006738

La probabilidad de que se registren cero accidentes en una calle en la que el promedio de accidentes diario es 5, es 0.006738.

Si deseamos saber la probabilidad de que ocurra exactamente 1 accidente ese día y en esa misma calle, repetimos el proceso, dando el valor de 1 = b

Traemos y llenamos los interrogantes PDF:POISSON(1,5)
El resultado, en la pantalla de datos, será

Ésa es la probabilidad de que ocurra exactamente un accidente en ese día determinado en la calle cuyo promedio diario es de 5 accidentes.

Repetimos el proceso para los demás casos.

7

MUESTREO Y DISTRIBUCIONES DE MUESTRAS

Introducción

Por lo general, no se tiene datos sobre las poblaciones totales a las que se desea analizar, por eso las muestras son tan importantes en la disciplina estadística.

Las muestras se realizan por medio de encuestas adecuadamente diseñadas y, aunque constiuyen sólo una porción del total de la *Población*, reflejan sus características.

Población

Es el conjunto total de elementos que conforman el universo sobre el cual realizaremos los análisis, generalmente sobre la base de muestras

Muestra

Es una porción escogida de la población y, cuando está adecuadamente diseñada, refleja las características de las mismas.

Tipos de muestreo:

Las muestras pueden ser aleatorias o no aleatorias; en las primeras todos los elementos de la población tienen la misma probabilidad de ser escogidos para la muestra.

Muestreo aleatorio

En estos muestreos conocemos las probabilidades de que un elemento de la población sea escogido en la muestra; se divide en las siguientes clasificaciones.

Muestreo Aleatorio Simple

Para comprender este concepto, diremos que se puede derivar, no una, sino muchas muestras de una sola población, una tras otra.

El muestreo aleatorio es el que establece que cada posible muestra tenga la misma probabilidad de ser seleccionada que cualquiera de las otras de la misma población.

También exige que cada elemento de la Población tenga la misma probabilidad que las demás de ser incluida en cualquiera de las muestras.

Muestreo estratificado

Se divide la población en grupos homogéneos, se pondera cada muestra, de acuerdo con el número de elementos que contiene con relación al total de las muestras.

La estratificación por barrios o por manzanos de barrio es una manera muy utilizada.

Fases del diseño de muestra

Determinar qué es lo que se desea medir.

Definir el tamaño de la muestra

Realizar un pequeño plan piloto con las hojas de encuestas en la que se anotan las variables a ser analizadas.

Introducción a las distribuciones de muestreo

Se pueden sacar muchas muestras diferentes de una población.

Supongamos que de una población extraemos varias muestras, una después de la otra, de manera tal que después de cada muestra los elementos vuelven a la población.

Lo mismo con las demás; supongamos que se trata de una muestra para estimar la altura de los varones tomados al azar, de una población de 100000 personas.

En este caso, la población puede ser considerada infinita; supongamos también que sacamos varias muestras del mismo tamaño de esa población.

Al comparar las muestras que se tomó para estimar la altura media y la desviación estándar, nos damos cuenta de que esos estadísticos varían de una muestra a otra.

Si los estadísticos varían de una muestra a otra, nos interesará saber cómo están distribuidas las muestras, cuando, teóricamente, tomamos todas las muestras posibles.

Es decir, si cada muestra que logramos de una población es diferente de las otras, también logradas de la misma población, cada una tendrá una media aritmética propia

Lo que ahora deseamos saber es cómo se distribuyen esas medias aritméticas

Vimos cómo el ingreso de las personas se distribuyen alrededor de la media aritmética, así también las medias se distribuyen alrededor de una media aritmética central.

Los matemáticos que se han encargado de formalizar las fórmulas y sistematizar los análisis han llegado a la siguiente conclusión:

> *Una distribución de probabilidad de todas las medias posibles de las muestras es una distribución de las Medias de la muestra.*

Los valores estadísticos se conocen como *distribución de muestras de la media*

Es muy importante diferenciar estos estadísticos de los anteriores.
A lo largo de este capítulo, aprenderemos a diferenciar ambos tipos de estadísticos.

Descripción de las distribuciones de muestreo

Ya lo vimos: cualquier distribución de probabilidad y, por lo tanto, cualquier distribución de muestreo, puede ser descrita por su media y por su desviación estándar.

Si sacáramos todas las muestras posibles de una población dada, las medias de cada muestra, al ser diferentes unas de las otras, estarían distribuidas de alguna manera.

Esa distribución de las medias tendría, a su vez, una media y una desviación estándar propios, dado que se trata de una distribución, aunque sean de las medias.

La Desviación Estándar que usábamos hasta ahora, toma otro nombre cuando se refiere a la distribución de las medias del conjunto de muestras: *Error Típico de la Muestra.*

Otros programas utilizan el nombre "Error estándar de la media", pero el SPSS designa a este estadístico como *error típico de la media.*

El Teorema del Límite Central

La media aritmética de la distribución de las medias aritméticas de las muestras será igual a la media aritmética de la población, sin importar el tamaño de las muestras

Por otra parte, los matemáticos han llegado a la siguiente conclusión, la que ahora conocemos como el Teorema del Límite Central:

> *A medida que el tamaño de las muestras crece, la distribución de las medias de las muestras se acerca a la distribución normal*

Por eso es que la muestra no tiene que ser muy grande con respecto a la población.

El Teorema del Límite Central nos permite usar los estadísticos de una muestra para hacer inferencias sobre los estadísticos de la población.

También se puede asegurar que en una muestra con más de 30 elementos, la distribución de su media empieza a acercarse a la distribución normal.

Por eso es que la primera de las acciones que realizamos al recibir una muestra en la pantalla del SPSS es constatar si su distribución es normal.

Si no es normal, posiblemente tengamos que transformar la muestra.
Existen varios procedimientos para ese objetivo.

Conclusión

Todo lo que hemos hecho hasta ahora ha sido recordar los principales conceptos que usa la estadística para analizar las muestras que logra de las diferentes poblaciones.

Excepto en casos, como los censos de población, vivienda, ingresos, todas las estimaciones sobre las estadísticas de una población se realizan utilizando la muestra.

Continuaremos nuestro estudio aplicando el SPSS.

La División principal de la Estadística

Lo dijimos ya, la Estadística se divide en dos grandes ramas:

La Estadística Descriptiva describe las propiedades de las muestras, incluyendo los valores, las tablas, las distribuciones, las gráficas y otros.

La Inferencia Estadística infiere las características de la población a partir del análisis de los resultados que nos brinda la Estadística Descriptiva.

La primera parte de este curso trata de la Estadística Descriptiva y la segunda parte estará orientada a la Inferencia Estadística.

Tipos de variables

A modo de recordatorio, repetiremos lo que se estableció en el primer capítulo.
Hay dos tipos principales de Variables: las categóricas y las de escala.

Variables categóricas

Son variables no numéricas, pero que pueden volverse numéricas en su caso.

También se las conoce con el nombre de variables cualitativas.
Las variables categóricas se dividen en nominales y ordinales.

Variables nominales son aquéllas que no necesitan de un orden preestablecido, tal el Estado Civil: Soltero, casado…

Variables Ordinales obedecen a un orden jerárquico de ordenación.

Marital status, por ejemplo, es una variable categórica nominal, pues no es cuantitativa, sobre todo, no necesita una jerarquía de ordenación.

Level of education es una variable ordinal, dado que sigue un orden establecido.

Ese orden está dado por la sucesión de los que no han asistido a un sistema formal de educación hasta los que han llegado a un nivel más allá de un grado universitario.

Variables de Escala

Son numéricos; también se llaman *Variables Cuantitativas.*

Frecuencias

Esta opción requiere de *Statistics Base option.*

El proceso *Frecuencias* nos ofrece el despliegue de estadísticos y gráficas que nos son necesarias para describir los tipos de variables que debemos interpretar.

Es un buen lugar para que empecemos a familiarizarnos con los datos.
Las gráficas pueden ser expresadas en valores absolutos o en porcentajes.

Usar la opción Frecuencias para analizar variables nominales

Traigamos el archivo *contacts.save* (Ver *Sample Files* para mayor información)

El archivo tiene información sobre un equipo de vendedores de computadoras para las empresas que desarrollan software.

En cada compañía, cada departamento dispone de representantes que tienen contacto primario; los contactos están categorizados con las siguientes denominaciones:

Development, Computer Services, Finance, Other, Don't Know

Se desea comprobar si los departamentos cumplen con los objetivos.
Usaremos *Frecuencies* para analizar la distribución de los departamentos.

*Analize → Descriptive Statistics → Frequencies → Departament (*como la unidad de análisis*) → Charts → pie chart → Continue → OK*

Tabla 7.1

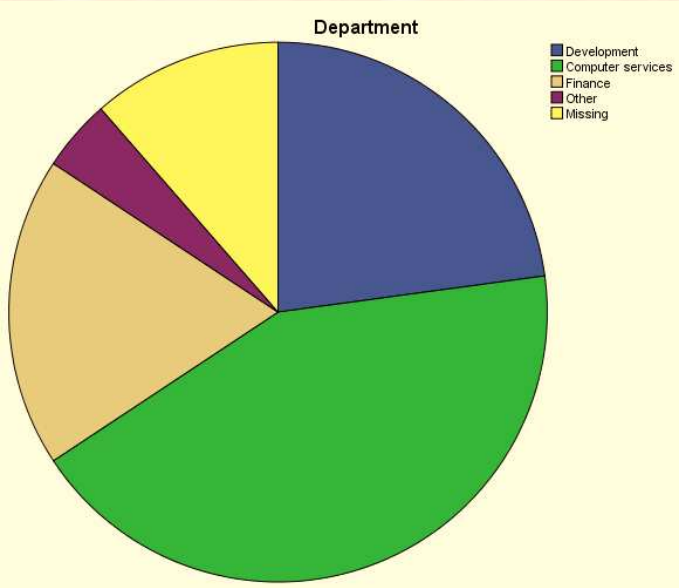
Department					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Development	16	22,9	25,8	25,8
	Computer services	30	42,9	48,4	74,2
	Finance	13	18,6	21,0	95,2
	Other	3	4,3	4,8	100,0
	Total	62	88,6	100,0	
Perdidos	Don't know	8	11,4		
Total		70	100,0		

En la pantalla de resultados tenemos una tabla y un gráfico de "torta".
Copiamos la tabla.

El total de observaciones registradas es 70, de las cuales 8 son valores perdidos, es decir, no incluidos en la muestra, lo que es equivalente al 11.4% del total.

La columna de frecuencias nos informa que 30 de los contactos personales se realizaron con el departamento de servicios de computadora.

Gráfica 7.1



Esto equivale al 42.9% del total de contactos y al 48.4% de los contactos del número de variables efectivamente tomadas en cuenta.

La información que leemos en las tablas numéricas se refleja en gráficos; la "Torta" es una ayuda para tener una idea de las frecuencias relativas de cada categoría.

Con las instrucciones anteriores, lo que hicimos fue concretar una tabla de datos.
La gráfica de "torta" se nos dio cuando hicimos click en Charts → Pie charts.

Copiamos el "Pie" para apreciar los valores relativos de una forma más visual

En la parte derecha de la gráfica 7.1 se registra los sectores que tienen contactos con empresas clientes de la firma que ahora estamos analizando.

Los respectivos porcentajes se muestran en las porciones de color de la torta.

Gráfica de barras

Para obtener una nueva clase de gráfico, vayamos a la caja de Frecuencias, pulsemos en *Charts →Bar charts → Continue → OK*

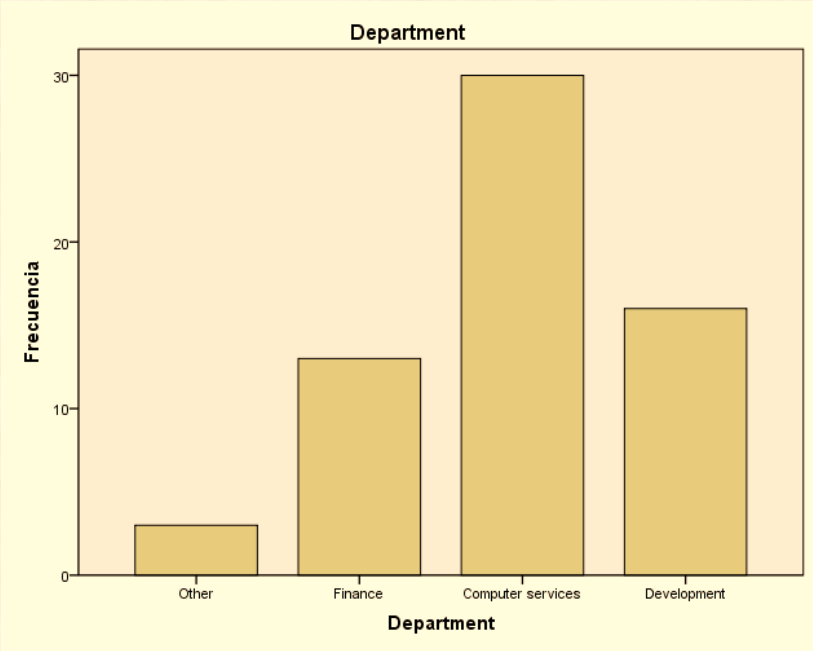
En el cuadro de frecuencias→ *Format → descending values→ continue → OK*
Para una mejor visualización, las barras se muestran de mayor a menor.

Además del departamento respectivo para cada contacto, también vemos el rango de las observaciones; para ello vamos otra vez a Frecuencias.

Cada una de las barras representa el número de clientes que tuvieron contacto con la empresa a través de sus departamentos respectivos: finanzas, desarrollo, etc.

Deseamos ver los rangos de la compañía para constatar si cumple con las metas.

Gráfica 7.2



Resumen de rangos

Analize→ Descriptive Statistics → Frequencies → Reset
En la pantalla de la derecha Introducimos *Company Rank* , que es variable ordinal

Gráficos → Charts → Bar Charts → Continuar → Format → Ascending Values → Continuar → OK en la caja de diálogo de frecuencias.

En la pantalla de resultados tenemos la tabla 7.2.
Al igual que las anteriores, la copiamos para su respectivo análisis.

Los resultados nos muestran que las variables ordinales también pueden ser analizadas de la misma manera y darnos información adecuada.

Por ejemplo, podemos ver en la tabla que los contactos con los administradores jóvenes constituyen el 15.7% del total de los contactos.

Pero, cuando estudiamos datos de tipo ordinal, el porcentaje acumulativo es más útil.

La tabla, al igual que las barras, ha sido ordenada por magnitudes descendentes.
Para esta clase de análisis, la variable a estudiarse tiene que ser cuantitativa.

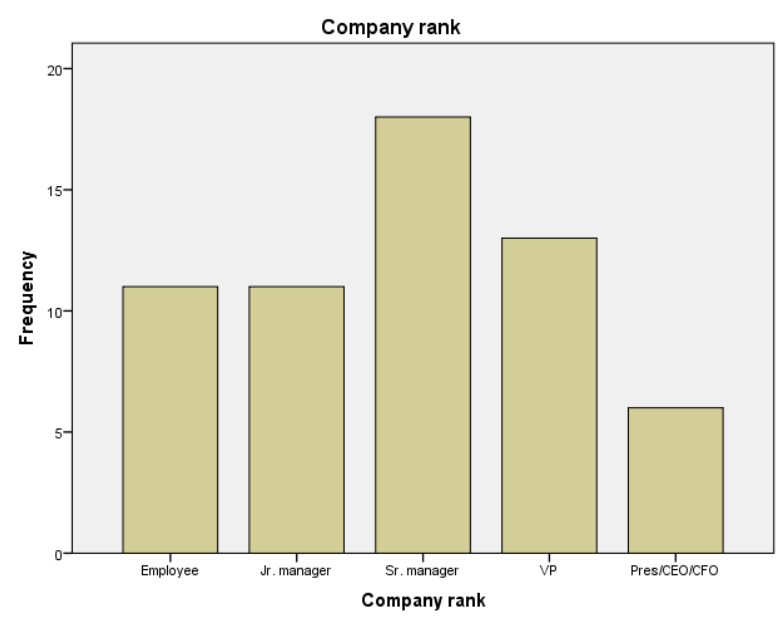
Tabla 7.2

Company rank					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Employee	11	15,7	18,6	18,6
	Jr. manager	11	15,7	18,6	37,3
	Sr. manager	18	25,7	30,5	67,8
	VP	13	18,6	22,0	89,8
	Pres/CEO/CFO	6	8,6	10,2	100,0
	Total	59	84,3	100,0	
Missing	Don't know	11	15,7		
Total		70	100,0		

Allí podemos observar que el 67.2% de los contactos tuvieron alcanzaron, desde employee a Senior manager, pasando por Jr. manager

La Gráfica 7.3 es el Histograma de la distribución de la variable Company Rank.
En el eje base registra las cinco categorías que se muestran en la tabla 7.2.

Gráfica 7.3



La Altura de cada barra del histograma consigna la frecuencia de cada categoría, es decir, la misma información que tenemos en la columna "Frecuencia" de la tabla 7.2

Frecuencias

Para analizar esta opción usamos la distribución de la variable *Amout of Last Sale, que se encuentra en el archivo contacts.save,* que estamos utilizando

Analize → DescriptiveStatístis → Frequencies
En la pantalla de la derecha aún está la variable Department

Para llevarla a la pantalla de la izquierda, pulsamos en la flecha de dirección y la trasladmos; allí click → Amount of Last Sale

Ahora vamos a llamar a varios estadísticos cuyo significado aún no conocemos
Sin embargo, los iremos analizando uno por uno, cuando los tengamos en las tablas.

Una vez en la caja de diálogo de frecuencias:

→ Statistics → click → *Quartiles → Std. Deviation → Mínimun → Máximun → Mean → Median → Mode → Skewness → Kurtosis → Continue* → Charts → *Histogram → Continue → OK*

*D*esactivamos la opción "tablas de frecuencias" → *Aceptamos*

Tabla 7.3

Statistics		
Amount of last sale		
N	Valid	70
	Missing	0
Mean		55,4500
Median		24,0000
Mode		8,00[a]
Std. Deviation		103,93940
Skewness		5,325
Std. Error of Skewness		,287
Kurtosis		34,292
Std. Error of Kurtosis		,566
Minimum		6,00
Maximum		776,50
Percentiles	25	12,0000
	50	24,0000
	75	52,8750
The smallest Mode is shown		

La *Tabla 7.3* nos muestra los estadísticos acerca de la distribución de la variable que analizamos, en este caso, de la variable "ventas"; definamos cada uno de ellos.

El título del cuadro, *Amount of Sales*, se refiere al total de lo que se vendió el último día; N = 70 significa que se tomó en cuenta 70 valores de la muestra.

No hay valores perdidos (Conocidos como "valores missing").

La Media nos da el promedio de las ventas últimas en todos los departamentos de la firma; en este caso, nos dice que el promedio (Mean) fue de $55.45.

La Mediana (Median), tal como ya la definimos, es de $24; de inmediato notamos la diferencia que hay con la media, lo que analizaremos en unos momentos

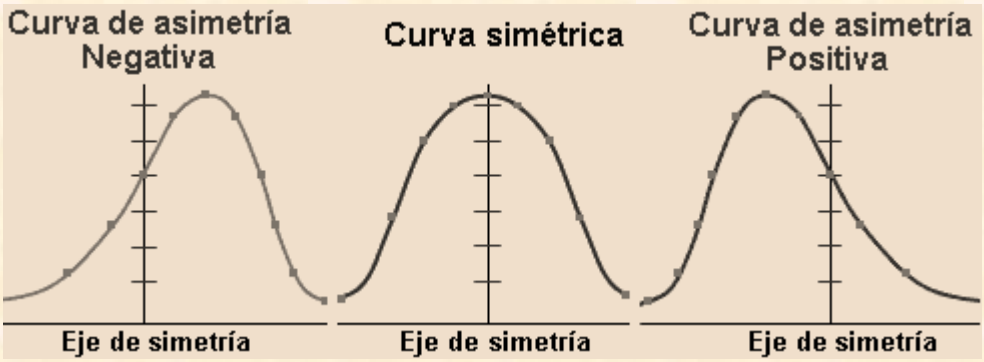
La Desviación Estándar (Std. Deviation), que muestra la dispersión media de las ventas, es $103.9394

Asimetría (Skewness o Sesgo)
Es un estadístico que describe la simetría de la distribución alrededor de la media.
Si el sesgo es igual a cero, la distribución es simétrica (en este caso, es normal)

Los siguientes gráficos ilustran el concepto de Asimetría.
Si no existe una cola o sesgo, la asimetría tendrá un valor de cero.

Si el sesgo lleva el signo positivo, entonces sabremos que la distribución tendrá una cola asimétrica hacia los valores negativos

Gráfica 7.4



Esto querrá decir que los elementos de la muestra en general estarán sesgados hacia los valores que se agrupan más en los niveles bajos.

Si el sesgo es negativo, la distribución tendrá una cola asimétrica hacia valores positivos; esto es, tiende a reunirse más a la derecha de la media, en los valores altos.

La curva de una distribución asimétrica con skewness (sesgo) negativa indica que los valores estarán más reunidos en niveles superiores a la media aritmética.

Al centro, la curva muestra la Curva Normal, la asimetría tiene un valor de cero, esto nos muestra que la distribución no está sesgada ni positiva ni negativamente.

A la derecha se muestra una curva asimétrica con skewness positiva, lo que nos indica que los valores tienden a reunirse en la zona de los valores menores a la media.

El valor de Asimetría que nos da la tabla 7.3 para nuestro análisis sobre las ventas últimas, es de 5.325; la Desviación Estándar (Std. Deviation) de la Asimetría es 0.287.

Es un valor positivo, por lo tanto las ventas en los departamentos, se reúnen en valores menores a la media, mostrando un desempeño no eficiente.

Curtosis

Es un estadístico que describe el grado de "punta" o "achatamiento" de la distribución de una variable con relación a la distribución normal.

La *gráfica 7.5* nos muestra el grado de curtosis de varias curvas.

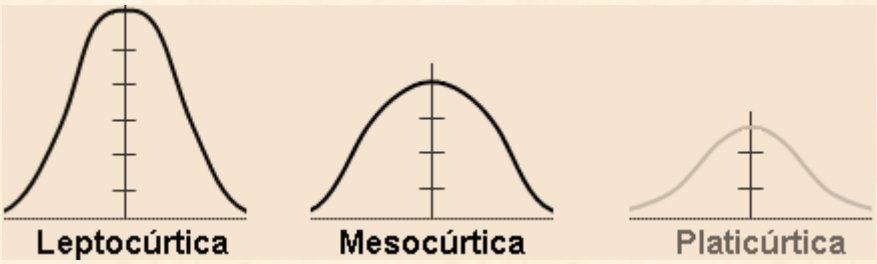
La curtosis positiva, llamada Leptocúrtica, indica una distribución que perfila un gráfico "más en punta", tal como vemos en la curva izquierda, con relación a la normal.

Una curtosis negativa, denominada Platicúrtica, indica una distribución relativamente más achatada, como la que vemos a la derecha de la gráfica 7.5

En una distribución normal la curtosis es igual a 3.

Es sobre ese valor que se determina el nivel de Curtosis de una distribución
En esta distribución, lo dijimos, la media, la mediana y la moda son iguales entre sí.

Gráfica 7.5



Si el coeficiente es positivo, quiere decir que hay una mayor concentración de los datos en torno a la media.

Si el coeficiente es negativo, la distribución Platicúrtica muestra que hay una menor concentración de datos en torno a la media; sería más achatada que la primera.

El valor de la curtosis en la tabla 7.3 es 34.2920, lo que indica que los valores están concentrados lejos de la media aritmética.

El uso del coeficiente de la Curtosis es muy importante para establecer el grado en el que la distribución de los valores se acerca o aleja de la curva normal.

Percentiles

Muestran cómo se agrupan los datos cuando se los ordena en orden ascendente tendiendo a que cada grupo tenga P% de observaciones.

Por ejemplo, podemos dividir el número de elementos de una muestra en 99% grupos a los que identificamos con los siguientes símbolos: P_1, P_2,.. P_{99}.

Cada grupo recibirá el nombre de percentil.
El percentil P_{100} ya queda establecido.

Estableceos que el 1% de los datos tienen un valor menor a P_1 (No olvidemos que los grupos se forman de una serie de datos ordenados en sentido ascendente)

El 70% de los datos serán menores o iguales a P_{70}.
El 99% de los datos serán menores que P_{99}.

Vamos a poner un ejemplo concreto, en el próximo subtítulo relativo a los Cuartiles.

Los Cuartiles

Son los grupos más utilizados, dividen al total de los elementos en cuatro partes cada uno delos cuales tienen aproximadamente el mismo número de elementos.

Cada Cuartil queda identificado como: Q_1, Q_2, y Q_3

Q_1, es el que tiene el primer 25% de observaciones del total y son menores a Q_1
Q_2 es el que tiene el 50% de las observaciones, cada uno es menor que Q_2
Q_3 es el que registra el 75% de los datos menores que Q_3

Ejemplo

Tomemos los siguientes valores ordenados en forma ascendente:

120, 140, 200, 240, 260, 380,450, 500, 630, 700, 750, 790, 800, 820, 910, 980

Ahora bien, dividamos ese total de elementos en cuatro grupos, es decir en Cuartiles

120, 140, 200, 240, 260, 380, 450, 500, 630, 700, 750, 790, 800, 820, 910, 980

El primer Cuartil (Q_1) es 240; todos los valores a su izquierda serán menores que él.
El Segundo Cuartil (Q_2) es 500; todos los valores a su izquierda son menores que él.

Del mismo modo para los demás cuartiles
El último Cuartil (Q_4) es 980, todos los elementos tienen un valor menor que 980

Obsérvese que la Mediana, tal como la estudiamos en los primeros capítulos, oscila entre 500 y 600, = 550, es decir, está incluida en el Segundo Cuartil

Es importante saber que el 50% de los datos centrales se encuentra entre Q_1 y Q_3

Los Deciles

Dividen la distribución en 10 partes iguales (9 divisiones) y se representan como D_1 ,..., D_9, correspondientes al 10%,...,90%, al igual que los otros estadísticos descritos.

El SPSS estima todos estos estadísticos

Volviendo al análisis de la tabla 7.3, notamos que la media, la mediana y la moda difieren en sus respectivos valores; la distribución de la variable no es normal.

La deducción se confirma al leer los estadísticos como la Asimetría y la Curtosis

El Histograma

Ya tuvimos ocasión de familiarizarnos con el Histograma, que es una visualización gráfica de la distribución de la variable que deseamos analizar.
Sin embargo, ahora necesitamos un Histograma que también incluya la curva normal, de esta manera podremos ratificar los datos de Kurtosis y de Sqweness

Analize → Descriptive Statistics → Frequency → Charts → Histograms → show normal curve on histogram → Continue → OK.

Obsérvese que en el eje de las abscisas de la Gráfica 7.6, se registran los valores de las ventas y en el de las ordenadas la frecuencia en que esos valores se repiten.

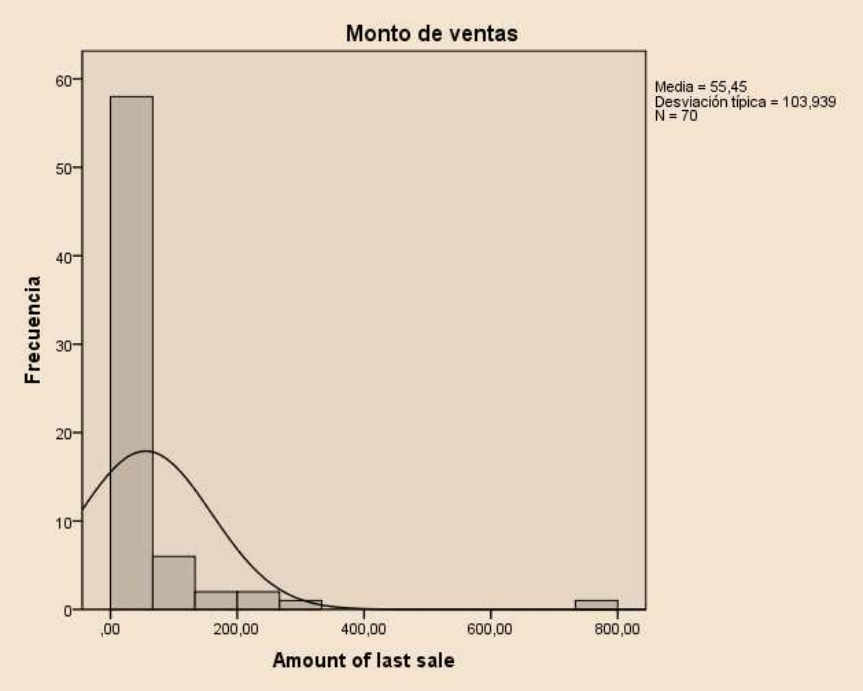
El histograma que vemos en el cuadro de resultados, comparado con la curva normal, nos muestra una distribución asimétrica con una cola derecha larga.

Es decir, un skewness positivo, lo que nos indica que los elementos están concentrados, en la zona de valores bajos, tal como nos indicaba la Asimetría.

También nos confirma una Curtosis elongada hacia arriba, con relación a la curva normal; la distribución es Leptocúrtica, tal como vimos en la tabla 7.3.

Los valores que nos da la tabla sobre las medidas de tendencia central, la *Curtosis y la Asimetría* muestran que la distribución de la variable *Sales* no es normal.

Gráfica 7.6



El Histograma confirma la misma conclusión que los datos de la tabla; por lo tanto, ya estamos seguros de que la distribución de la variable "sales", ventas, no es normal.

Deseamos que la distribución de la variable se acerque a la curva normal.

Para ello, recurrimos a la transformación de las variables de la muestra en otras que reflejen mejor las características.

Es decir, trataremos de transformar la variable de manera tal que su distribución se acerque más a la distribución normal

Transformación de variables

En el siguiente paso veremos cómo se hace esa transformación.
Primer, traemos a la pantalla de datos el archivo

La Opción Transformación

Para usar la opción Transformación, la variable debe ser de escala, cuantitativa.

En este caso, parece adecuado transformar los valores en los logaritmos naturales de los valores originales; la prueba y error con el SPSS nos enseñará a escoger.

Vamos a transformar la misma variable, *Amount of last sales* del archivo que estamos usando, esto es, de contacts.save

Por algunos segundos, sin hacer click, apuntamos con el cursor la columna *sale*, tendremos el nombre de la variable: *Amount of last sales*

Para mejorar la distribución de esta variable es que vamos a convertir sus valores originales en sus respectivos logaritmos

Para transformar la variable de ventas escojamos del Menú principal la casilla

Transform → C*ompute Variable*

Aparecerá un cuadro de diálogo

En realidad ya estamos familiarizados con esta opción, pues la usamos en el capítulo relativo a la solución de problemas de distribución de Probabilidades

En la parte superior izquierda del cuadro hay una casilla con el nombre "Target Variable", allí escribamos *logsale* que es *Amount of last sales* logaritmizada

A la derecha hay una pantalla con el título de *Numeric Expresion*

Ahora bien, debajo de esa pantalla, a la izquierda, hay dos pantallas: la primera dice "grupo de funciones"; de las opciones que nos ofrece pulsamos *All*.

La pantalla que se encuentra debajo de la anterior tiene como título *Functions and Special* Variables en la que aparecen todas las funciones del SPSS por orden alfabético.

Buscamos hacia abajo, hasta llegar a Ln; hacemos click y en la pantalla a la izquierda y debajo del teclado aparece una leyenda que explica cuál es la tarea de la función.

Con la flecha de dirección hacia arriba, introducimos esa función en la pantalla "expresión numérica"; allí aparecerá la siguiente expresión: LN(?)

El signo de interrogación pregunta cuál es la variable que deseamos logaritmizar

Del cuadro de la izquierda, donde están todas las variables elegimos *Amount of last sales*; con la flecha de dirección introducimos la variable entre los paréntesis; OK.

Vamos a la Pantalla Variable View; allí se ha creado una nueva variable *logsale.*
El número de decimales es 2, lo aumentamos a 6

Vamos a la pantalla de *Vista de datos* y nos encontramos que los valores de la nueva variable *logsale* están registrados en la última columna de la pantalla.

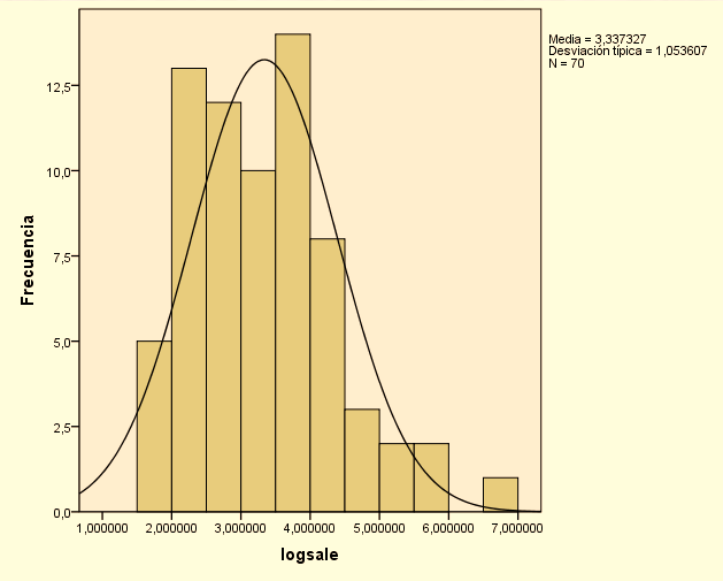
Observemos también que la nueva variable *logsale* ha quedado registrada en la pantalla *Variable View*; aumentamos los decimales a 6

En la columna *Etiquetas* ponemos *Logaritmo de Ventas.*

Ahora comprobaremos si la transformación de la variable sales ha mejorado con la transformación de sus valores en valores representados por sus logaritmos naturales.

Menú → Descriptive → Statistics → Frequencies

Gráfica 7.9



Desplazamos la variable sales de la pantalla y la reemplazamos por la nueva *logsale* → pulsamos los estadísticos deseados.
Charts →Histograma, la que trasladamos en la página anterior como la gráfica 7.9.

Aceptamos y una nueva tabla de datos y un nuevo histograma aparecen en la pantalla de resultado; analizamos primero el nuevo Histograma

Una revista al histograma confirma los datos numéricos de la tabla, pues el nuevo histograma se aproxima mucho más a la curva de la distribución normal.

Obsérvese que sobreponiéndose a las barras del histograma se muestra una curva normal, que es la que hemos solicitado para compararla con el histograma.

Otra vez podemos apreciar que la nueva distribución es más cercana a la normal.
Ahora traemos la tabla 7.4

Al leer los datos vemos que la media, la mediana y la moda están mucho más cerca entre sí, de lo que estaban antes de la transformación de la variable.

Notamos que los valores de C*urtosis y Asimetría* se han reducido, mostrando que la distribución de la nueva variable *logsale* se acerca más a la distribución normal

Tabla 7.4

Statistics		
logsale		
N	Valid	70
	Missing	0
Mean		3,33732686
Std. Error of Mean		,125930091
Median		3,16328809[a]
Mode		2,079442[b]
Std. Deviation		1,053606729
Variance		1,110
Skewness		,721
Std. Error of Skewness		,287
Kurtosis		,367
Std. Error of Kurtosis		,566
Range		4,863037
Minimum		1,791759
Maximum		6,654797
Percentiles	25	2,52416766[c]
	50	3,16328809
	75	3,96081317

Resumen

Hemos analizado las distribuciones de los contactos que se tiene con los encargados de las compañías para las que la firma produce hardware.

Se ha analizado la variable de ventas y sus valores han sido transformados en logaritmos naturales, para que su distribución se acerque más a la distribución normal.

Para ello, hemos utilizado la Opción *Transform*, lo que ha permitido que los valores con los que se debe trabajar en adelante sean mucho más útiles que los anteriores.

Al final del ejercicio sacaremos los antilogaritmos de los valores pronosticados, pero eso será cuando lleguemos al capítulo de Regresión.

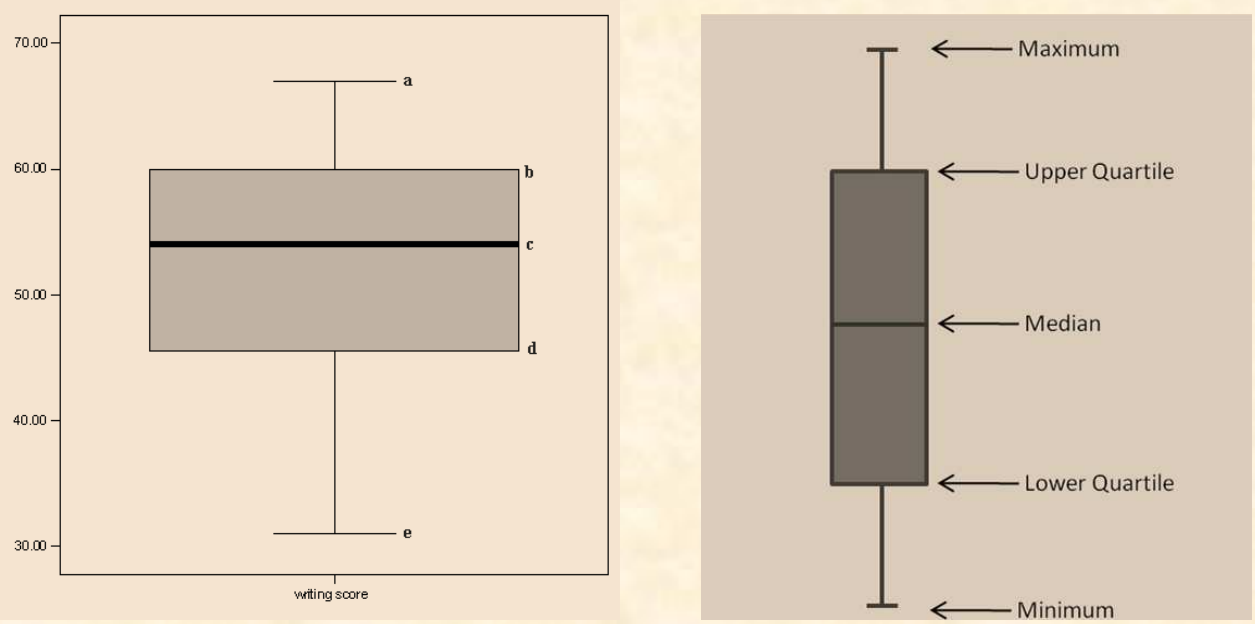
Diagrama de Caja (Boxplots)

El Diagrama de Caja muestra indicadores sobre el grado de normalidad de la distribución de la variable; especialmente nos muestra la asimetría y la curtosis.

En este capítulo, aprenderemos cuále son sus características y en el capítulo 8 tendremos casos concretos; en la *gráfica 7.4* tenemos un diagrama de caja típico.

La caja está entre dos verticales que apuntan hacia arriba y hacia abajo respectivamente; sus longitudes muestran el grado de asimetría de la distribución.

Gráfica 7.7
Diagrama de Caja



Cada una de las líneas verticales, hacia arriba o hacia abajo, se llama *Whisker.*

Si la distribución de los datos de la muestra fuera normal, ambos *whiskers* tendrían la misma longitud, pero eso sucede muy pocas veces.

Si el *whisker* superior es más corto, la muestra tendrá una asimetría negativa, los elementos estarán más concentrados en los valores superiores a la Media.

Si el *whisker* inferior es más corto, la asimetría será positiva, los elementos de la muestra estarán preferentemente concentrados en los valores inferiores

Cada whisker termina en una línea horizontal, la que marca sus límites.
Las Cajas también nos proveen información de la c*urtosis* de cada distribución.

Una caja delgada muestra que una gran cantidad de valores están contenidos en un segmento muy pequeño de la muestra; una distribución con un pico más alargado

Una caja inusualmente ancha significará lo contrario; los *outliers*, o valores extremos, están simbolizados por puntos, círculos o asteriscos más allá de los límites.

Los valores extremos se desvían por encima o por debajo de los otros valores.

La longitud de la caja nos ofrece un referente acerca de la variabilidad de la distribución de la muestra.

En la gráfica izquierda de 7.4 vemos letras que representan las características de la misma.

El valor máximo dentro de una distancia que no supere 1.5 veces el rango intercuartílico, el que ya sabemos, es Q_3, está representado por la letra **a.**

Ese espacio muestra el tercer cuartil más 1.5 veces el rango intercuartílico.

b es el tercer cuartil (Q_3), también es el 75 percentil
c es la mediana (Q_2) o también el porcentil 50

d es el primer cuartil (Q_1) o también es el percentil 25.
e es el valor mínimo.

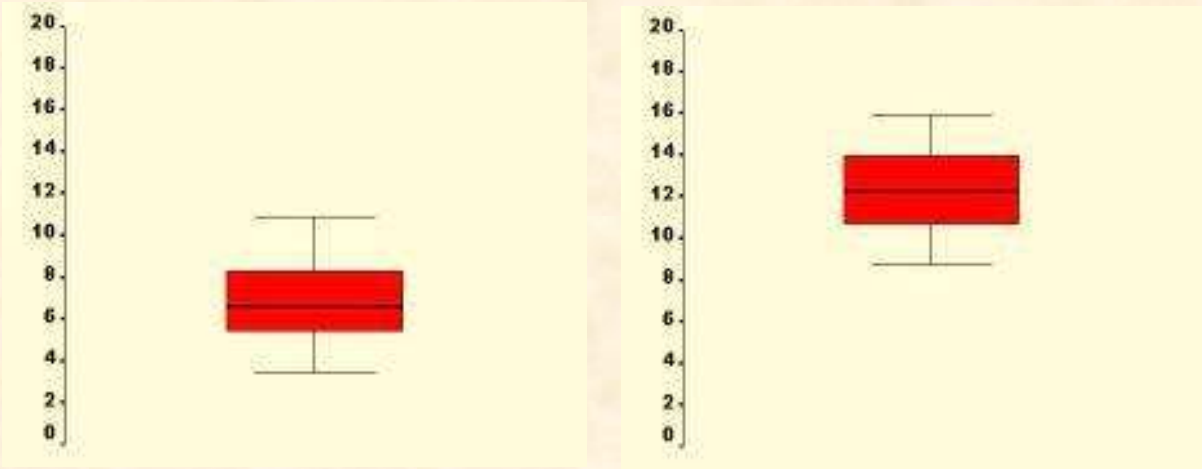
Pero si hay valores menores a 1.5 veces el rango intercuartílico debajo de Q_1, será el primer cuartil menos 1.5 veces el rango intercuartil.

El estadístico Diagrama de Caja o Boxplot es uno de los más complejos para captar a primera vista la estructura de la distribución.

Por eso es que analizaremos con mayor detalle sus características, las que nos permitirán conocer la estructura de la distribución de los elementos de una muestra.

La Caja derecha del gráfico anterior, muestra los puntos importantes de un Boxplot

Gráfica 7.5
Diagramas de Caja mostrando centralidad



Al analizar una Caja de Diagrama debemos recordar siempre que la línea del centro es la Mediana de la distribución que deseamos analizar.

La Caja de la izquierda de la gráfica 7.5 representa una muestra de 20 valores con una mediana centrada en 7 (representada por la línea gruesa interna)

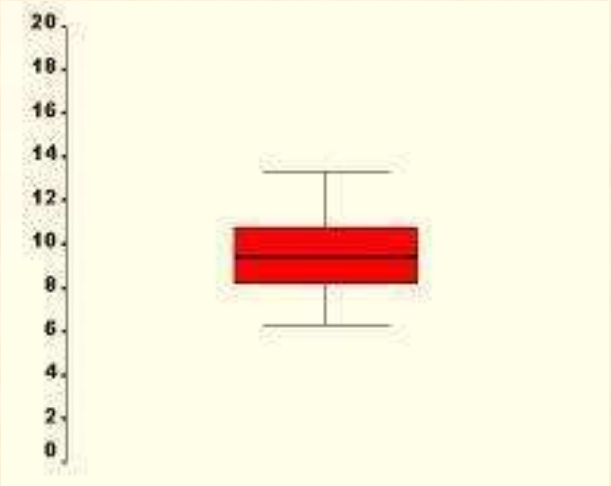
La Caja de la derecha: una muestra de 20 valores cuya mediana está centrada en 12.

La altura se hace más grande cuando la desviación estándar es mayor, como en el panel derech de la gráfica 7.5

Con las figuras (traídos de Internet, aunque no se pudo identificar al autor) las características y utilidad de la Caja de Diagrama se hacen cada vez más patentes.

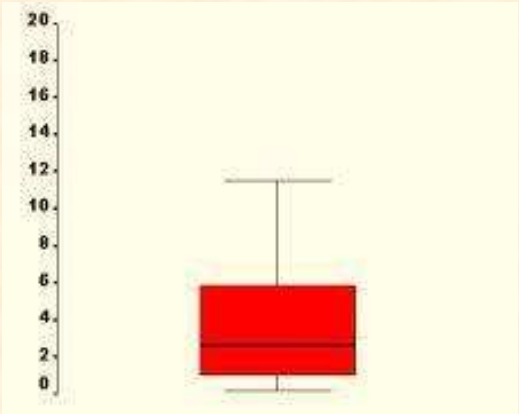
En la siguiente página veremos diferentes cajas de Diagrama o boxplots como Indicadores de Curtosis, tal como se muestra en la gráfica 7.7

Gráfica 7.7



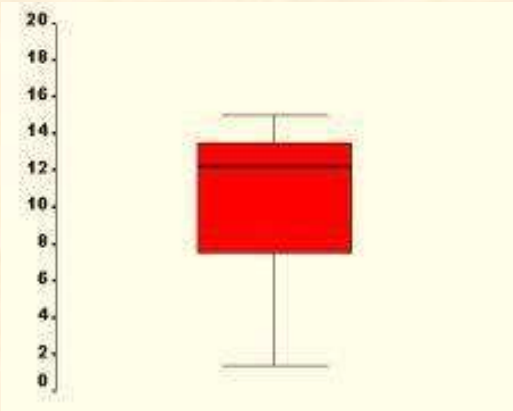
En esta caja, los 20 valores de la muestra tienen una distribución mesocúrtica.

La línea gruesa, esto es, la Mediana, está cerca del centro y los whiskers son de igual longitud; se acerca a la Normal



En esta caja, también de 20 valores, la curva es achatada, platicúrtica, pues su mediana está en la parte inferior.

La distribución de 20 valores es lepocúrtica
La Mediana está en la parte superior de la Caja



Se denominan hinges *a los valores* que están incluidos en la caja misma.

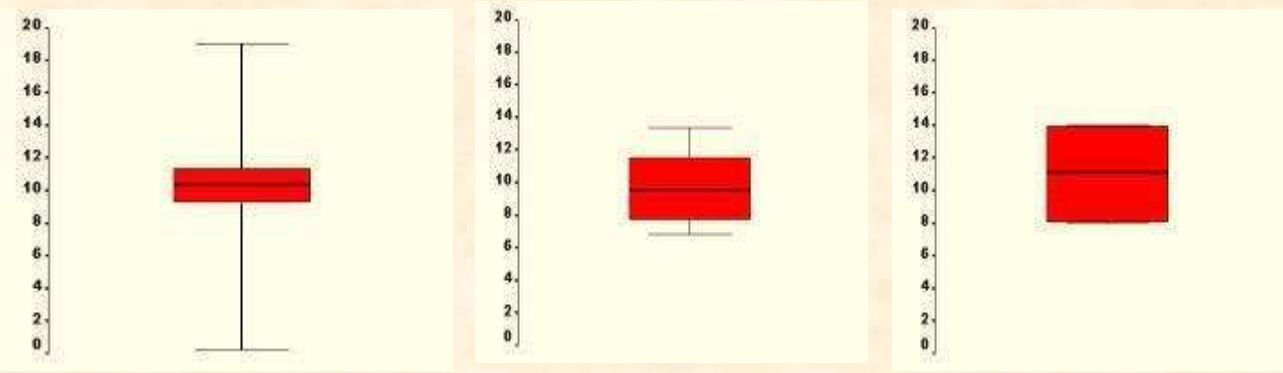
La Caja de Diagrama como Indicador de Asimetría

Las colas son los dos extremos, izquierdo y derecho, de una curva de distribución; en la curva normal, esas colas son asintóticas con relación al eje de las abscisas.

La Asimetría se muestra a través de la diferencia de longitudes entre las colas, pues, si la distribución es asimétrica, una de ellas será más larga que la otra.

La Curtosis ideal sería simétrica, como es la de la curva normal, cuyo diagrama de caja mostrará sus dos whiskers casi de la misma longitud.

Gráficas 7.8
Representación de la Asimetría



Las cajas que se diseñan en la gráfica 7.8 muestran las formas de asimetría de la distribución que representan en cada caso.

En la caja de la izquierda, la muestra de 20 valores tiene una distribución de largas colas, lo que se expresa en la longitud de los *whiskers*.

Cada longitud es mucho mayor que la altura de la caja.

Si la distribución se acercara a la normal, la longitud de cada *whisker* tendería a ser igual a la altura de la caja, como el bolxplot del centro

En el segundo ejemplo, la longitud de los *whiskers* es menor a la altura de la caja, lo que nos indica que la curtosis es menor que en la caja de la izquierda.

En la tercera figura, la distribución de la muestra tiene colas muy cortas y los *whiskers* están ausentes, esto es, la distribución que representa se acerca más a la normal.

Interpretación

Mostradas la utilidad de la caja de diagramas, es preciso recordar que se debe evitar deducir conclusiones cuando las muestras son muy pequeñas

Por ejemplo, 10 valores o menos

Por otra parte, es necesario tomar en cuenta los valores que no están en la caja ni en la longitud de los *whiskers*, que son los que son valores extremos.

Así, los *whiskers* realmente no muestran los valores mínimos y máximos de una muestra, sino los que están a una distancia "razonable" de la caja, abajo o arriba.

Una muestra con gran asimetría y *whiskers*, con muchos valores inusuales al costado puede aparecer razonablemente simétrica en la caja.

Pero el SPSS tiene sus controles sobre estos casos; los valores que son más de tres veces la longitud de la caja, arriba o abajo, reciben una anotación roja.

Esos valores son identificados con asterisco y entran como valores extremos

Los valores que tienen entre 1.5 hasta 3 veces la longitud de la caja, arriba o abajo, tienen una anotación amarilla; son considerados como *outliers.*

Debemos recordar que las palabras *outliers* y *extremes* deben ser interpretadas en relación a la distribución de una curva normal.

En todo caso, debemos tener en cuenta que los valores extremos son valores que están por arriba o por debajo de los valores *outliers,* no son sinónimos.

8

PROBLEMAS ADICIONALES EN ESTADÍSTICA DESCRIPTIVA

Datos Cuantitativos

Este capítulo será dedicado a identificar los gastos que los consumidores hacen en el rubro de telecomunicaciones, como un ejercicio para el uso del SPSS.

El objetivo es la identificación de cuál de los servicios rinde mayor beneficio en una empresa determinada.

Para ello, tomaremos los datos de una compañía virtual de telecomunicaciones, la que tiene una base de datos de los consumidores que incluye, entre otros:

Información acerca de cuánto gasta un cliente en llamadas a larga distancia, llamadas libres, alquiler de equipo, carta de llamadas y servicio

La información pertinente está en el archivo *telco.sav* del SPSS
Lo traemos a pantalla para analizar los datos por medio de la opción *Descriptive.*

→Menú → Analize → > Descriptive Statistics → Descriptive
→Long distance last month → Toll free last month→ Equipment last month,
→ Calling card last month → Wireless last month → Options

→Mean →Std. Deviation → Variance → →Kurtosis → →Skewness → Continue → OK

La tabla que aparece en el cuadro de resultados muestra los estadísticos sobre el gasto, los que serán comparados con los gastos en cada uno de los servicios.

Pero es difícil distinguir en la tabla cuáles servicios son más rentables.

En promedio, los clientes gastan más en *equipment rental*, pero hay una gran escala de variación en las sumas gastadas.

Los clientes de *calling card service* gastan un poco menos en promedio, que en *equipment rental,* pero hay menos variación en los valores que representan el gasto

Se puede encontrar clientes que gastan mucho más o mucho menos que otros clientes en cada servicio; eso lo conseguimos estandarizando los valores de las variables.

Tabla de Resultados 8.1

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Long distance last month	1000	11,723	10,363	2,966	,077	14,052	,155
Toll free last month	1000	13,274	16,902	1,784	,077	8,325	,155
Equipment last month	1000	14,219	19,068	,846	,077	-,728	,155
Calling card last month	1000	13,781	14,084	1,576	,077	4,612	,155
Long distance over tenure	1000	574,050	789,974	3,092	,077	14,216	,155
Valid N (listwise)	1000						

Estandarización de Variables

Las variables estandarizadas se conocen como "z scores."
Pero hay un problema.

Los valores *z scores* requieren que las distribuciones de las variables sean cercanas a la distribución normal, condición que las variables escogidas no cumplen.

Tal como vimos los valores de los estadísticos asimetría y kurtosis son altos, es decir, las distribuciones de las variables no son normales.

Un remedio posible, debido a que todos los valores de las variables son positivos, es transformar las variables en sus logaritmos naturales y luego, estudiar los z scores.

Recordemos el ejercicio anterior para logaritmizar las variables; en esta ocasión, el SPSS ya las ha logaritmizado tal como las vemos en la pantalla Data View

Trabajaremos con esos valores ya logaritmizados del archivo telco.sav

Menú → Descriptive Statistics → Descriptive → Options

Si en la pantalla de la derecha aún quedan las variables no logaritmizadas, con la flecha de dirección las trasladamos hasta el menú de la izquierda

Del cuadro a la izquierda inferior seleccionamos las siguientes variables:

→Log-long → Log-wireless → Log-Equipment → Log-Calling Card → Log-wireless
→Options → Mean → Std. Deviation →Kurtosis → Skewness → Continue →
→Save standardized values as variables → OK

Copiamos la Pantalla de Resultados

Tabla de Resultados 8.2

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Log-long distance	1000	2,1821	,73455	,166	,077	-,001	,155
Log-toll free	475	3,2397	,41381	,304	,112	1,107	,224
Log-equipment	386	3,5681	,27756	,037	,124	-,344	,248
Log-calling card	678	2,8542	,55729	,081	,094	,109	,187
Log-wireless	296	3,5983	,36729	,200	,142	-,168	,282
Valid N (listwise)	131						

Los valores logaritmizados aparecen en las nuevas columnas de la pantalla de Vista de datos; también aparecen en la pantalla Vista de variables.

En la pantalla de Resultados aparece la tabla de valores; vemos que la asimetría y la curtosis han disminuido con relación a los valores de las variables no logaritmizadas.

Sin embargo hay algo más que debemos hacer

Convertir a Valores Missing

Si nos fijamos en las nuevas variables, ya logaritmizadas, nos encontraremos que en muchas casillas que en vez de datos tienen puntos

Ésa es la razón por la que copiamos un segmento de la Pantalla Data View
Allí podemos visualizar que hay varias casillas vacías en las variables

Esto quiere decir, que no hubo la información necesaria para llenar las casillas vacías.
Estas casillas pueden ser llenadas con lo que se llama *Valores Missing*

Con esa operación tratamos de mejorar la calidad de la información respectiva

Tabla 8.1

loglong	logtoll	logequi	logcard	logwire
1,31	.	.	2,01	
1,48	3,03	.	2,72	3,58
2,90	2,89	.	3,41	
2,25	.	.	.	
1,84	.	.	.	
2,47	2,96	.	2,60	
2,39	.	.	2,17	
1,80	3,81	3,91	3,15	4,17
2,28	3,35	.	2,48	
3,18	.	.	2,80	
1,58	.	3,26	.	

La traducción de Missing es: Valores perdidos.
Nuestra tarea ahora consiste en tratar las casillas vacías con Valores Missing

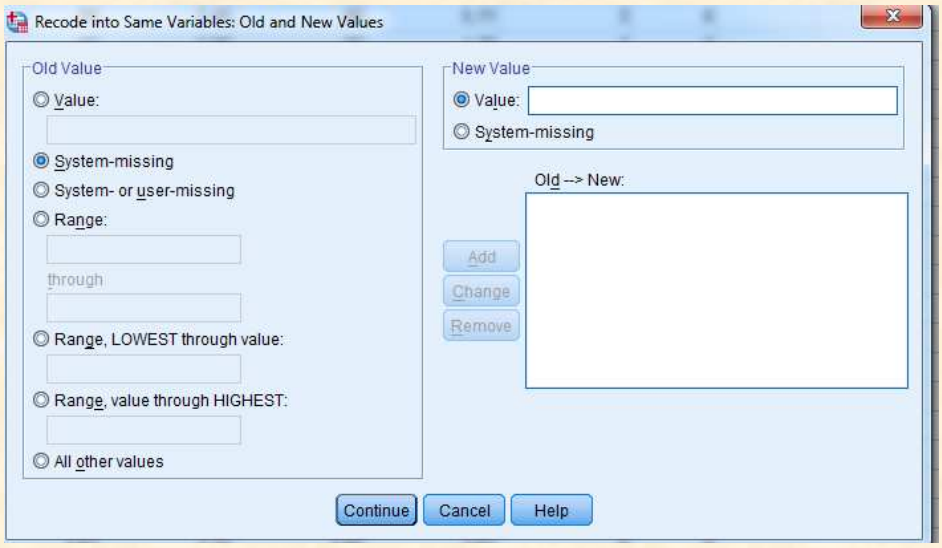
Menu → Transform →Recode into same variables

Aparece un menú con las variables a la izquierda y una pantalla vacía a la derecha

Con el mouse usamos la flecha de dirección para llevar las variables logaritmizadas desde la pantalla del menú a la pantalla derecha.

Debajo de la pantalla derecha hay una casilla que dice: *Old and New Values*
Hacemos click en esa opción y tendremos la siguiente caja

Cuadro 8.1



En la casilla superior izquierda de la sección primera del cuadro, pulsamos en el botón que dice Value

En la parte derecha también aparece otro botón Value, donde anotamos 0 y luego pulsamos la opción Add para traer el valor 0 a la pantalla inferior

En esa pantalla aparecerá la siguiente inscripción: SISMYS → 0
Continue → OK

En la pantalla Data View veos que en las variables que hemos logaritmizado, las casillas que tenían puntos, ahora tienen 0

Traemos un segmento de esa Pantalla

Tabla 8.2

loglong	logtoll	logequi	logcard	logwire
1,31	,00	,00	2,01	,00
1,48	3,03	,00	2,72	3,58
2,90	2,89	,00	3,41	,00
2,25	,00	,00	,00	,00
1,84	,00	,00	,00	,00
2,47	2,96	,00	2,60	,00
2,39	,00	,00	2,17	,00
1,80	3,81	3,91	3,15	4,17
2,28	3,35	,00	2,48	,00
3,18	,00	,00	2,80	,00
1,58	,00	3,26	,00	,00
1,96	3,09	,00	3,17	,00
2,15	,00	,00	3,73	,00
2,75	3,83	3,84	,00	4,11
1,48	,00	,00	,00	,00
1,63	,00	3,41	2,42	,00
2,78	3,39	3,44	3,40	,00
1,89	2,92	,00	,00	,00

Si comparamos la pantalla 8.2 con la anterior, 8.1 veremos que en las casillas que tenían puntos, ahora aparecen ceros

Más adelante veremos la utilidad de convertir los valores missing.
Sin embargo, hay algo que observamos en la Tabla de Resultados 8.2.

La variable *Log toll-free* tiene aún una kurtosis apreciable debido a que un cliente gastó una suma mensual considerable en el último mes.

Para analizar esta clase de valores inusuales debemos visualizar los "z scores".

The z-scores

Un z-score es una variable que ha sido estandardizada
Luego tendremos esa definición

Para trabajar con Diagramas de Caja, debemos estandardizar las variables

Menu → Analize → Descriptive Statistics → Descriptive

Con la flecha de dirección introducimos en la pantalla blanca de la derecha las cinco variables que han sido logaritmizadas: Desde *Log-long distance* hasta *Log-wireless*

Debajo del menú de la izquierda activamos el botón:

Save Standardize Values as Variables→ OK

En la Pantalla de Data View veremos inscritas nuestras variables logarítmicas precedidos por una Zeta; v.g Zlonglong, Zlogtall….

Con las variables así estandardizadas, podemos ir a trabajar con los Diagramas de Caja

Diagramas de Caja con el SPSS

Ya estamos familiarizados con las características del Boxplot o Diagramas de Caja

Traemos el archivo *Telco.save* a la pantalla de Vista de Datos.
Empezamos con llamar al Diagrama de Caja

Menu → Graphs → Legacy Dialogs → Boxplot → Simple →
→ Summaries of separate variables → Define

En la pantalla a la derecha introducimos las variables convertidas en z-scores

Options → Exclude cases variable by variable → Continue → OK

La Pantalla de Resultados registra una tabla y el Diagrama de Boxplots of Z-scores
Éste es el diagrama que necesitamos

La copiamos para traerla y analizar las características que tiene

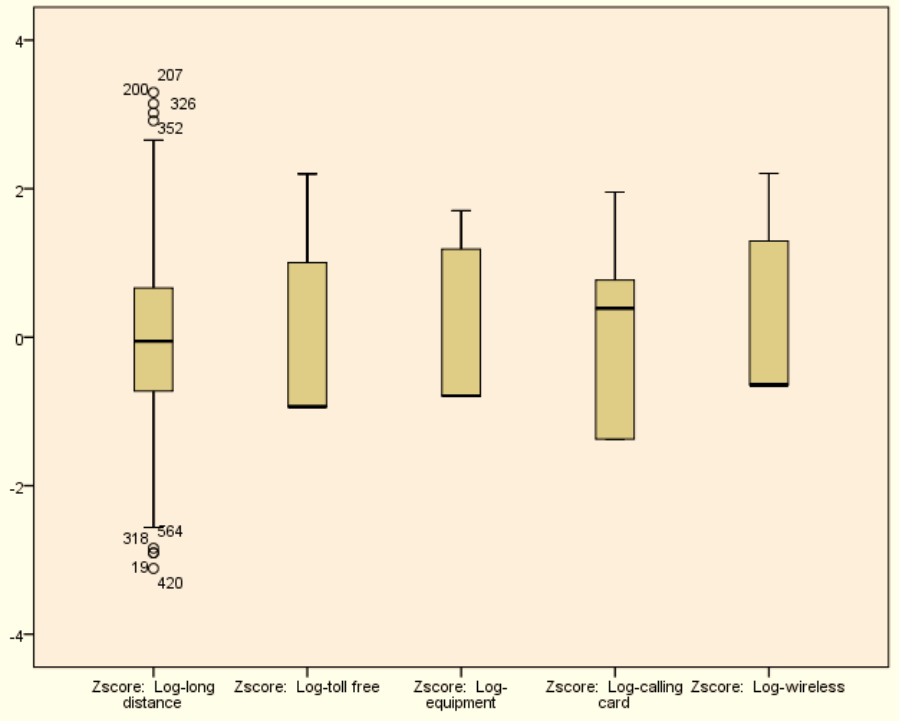
En los diagramas de caja vemos, por ejemplo, que en la variable Log-long distance hay valores extremos, 200 y 207

Estos valores son extremos, lo que definiremos de inmediato

Un outlier, lo vimos ya, es una observación cuyo valor varía significativamente de los valores de la mayoría de las observaciones.

Técnicamente se define como un valor cuya distancia al cuartil más próximo es superior a 1,5 veces el rango intercuartil.

Diagrama de Caja 8.1



Los outliers deforman la muestra porque tienden a sesgarla hacia valores extremos

Para tener una mejor visión de la información que nos proporcionan los Diagramas de Caja, vamos a llamar a otro archivo y a usar otra opción del SPSS

La Opción Explore y el Análisis de datos

La exploración de los datos, por medio del procedimiento *Explore,* ayuda a determinar si las técnicas de la ciencia estadística que usamos son o no apropiadas.

El procedimiento *Explore* nos provee de resúmenes visuales y números de los datos, ya sea para todos los casos por separado o ya para grupos de casos.

La variable dependiente a ser analizada debe ser una variable de escala.
Las de grupo pueden ser nominales u ordinales.

Con *Explore* podemos obtener los diagramas de caja, identificar los outliers, chequear los supuestos y las diferencias entre los grupos.

Estadísticas descriptivas cruzando grupos

Ejemplo; las cosechas de maíz deben ser testadas para establecer si hay o no el elemento aflatoxin, un veneno cuya concentración varía entre las cosechas.

Una firma que procesa granos ha recibido 8 diferentes cosechas, pero la distribución de aflatoxin, en partes por billón, debe ser estimada antes de que sean aceptadas.

Este ejemplo usa el archivo *aflatoxin.sav* cuyos datos consisten en 16 muestras de cada una de las entregas de las ocho en que se han transado la operación.

Menú → Analize → Descriptive Statistics → Explore

Aflatoxin PPB como la variable dependiente y *Corn Yield* como la variable factor → OK

En la pantalla de resultados tenemos el cuadro general para las 8 entregas (Yields) y las 16 muestras que se ha elegido de cada una.

Tabla de Resultados 8.3

	Corn Yield			Statistic	Std. Error
Aflatoxin PPB	1	Mean		20,2500	1,07819
		95% Confidence Interval for Mean	Lower Bound	17,9519	
			Upper Bound	22,5481	
		5% Trimmed Mean		20,4444	
		Median		21,5000	
		Variance		18,600	
		Std. Deviation		4,31277	
		Minimum		12,00	
		Maximum		25,00	
		Range		13,00	
		Interquartile Range		8,00	
		Skewness		-,788	,564
		Kurtosis		-,655	1,091
				33,0625	3,04339

Luego nos da la tabla con los estadísticos respectivos para cada una de las 8 entregas Los cuadros son muy extensos para copiarlos.

Pero traemos los datos para la primera entrega de las ocho, Tabla de Resultados 8.3
En la segunda Columna están registrados los valores de cada indicador.

En la tercera columna se registran los valores del error estándar, si corresponde.
En la tabla de resultados 8.3, hay tres nuevos conceptos que debemos captar

Interquartile Range

El Rango Intercuartil es la diferencia que hay entre el tercer y el primer cuartil de una distribución; es una medida de dispersión estadística: RI = $Q_3 - Q_1$

Si dividimos la expresión anterior entre dos, tendremos la desviación intercuartil
Esta medida es importante cuando la medida de tendencia central ha sido la Mediana

Se usa para estructurar y análizar los diagramas de Caja y bigote
Éste último será analizado en el último capítulo de la obra

Confidence Interval

El Intervalo de confianza de un indicador estadístico está conformado por dos valores límite: el límite inferior y el límite superior.

Es muy importante, pues nos ofrece mayor seguridad en nuestras apreciaciones.
Ejemplo, Supongamos que la Media de una distribución es 45

Con los instrumentos que aprenderemos en esta obra, podríamos establecer por ejemplo que la Media citada se encuentra entre los límites: 43 y 47

En este ejemplo, el límite inferior será 43 y el superior, 47
Esto quiere decir que la Media puede tomar algún valor entre esos límites

En la Tabla de resultados, la Media = 25

El límite inferior es 17,9519
El límite superior es 22,5481

Esto es, la Media podría tomar cualquier valor entre los límites señalados.
Esto se debe a que tratamos con muestras, no con poblaciones.

Luego veremos las probabilidades de que ése sea el Intervalo de Confianza

Trimmed Mean

Es una Media a la que se le ha recortado el %% de los valore extremos, tanto inferiores como superiores, para evitar sesgos inapropiados

La Opción Explore y el Análisis de datos

La exploración de los datos, por medio del procedimiento *Explore,* ayuda a determinar si las técnicas de la ciencia estadística que usamos son o no apropiadas.

El procedimiento *Explore* nos provee de resúmenes visuales y números de los datos, ya sea para todos los casos por separado o ya para grupos de casos.

La variable dependiente a ser analizada debe ser una variable de escala.
Las de grupo pueden ser nominales u ordinales.

Con *Explore* podemos obtener los diagramas de caja, identificar los outliers, chequear los supuestos y las diferencias entre los grupos.

Estadísticas descriptivas cruzando grupos

Ejemplo; las cosechas de maíz deben ser testadas para establecer si hay o no el elemento aflatoxin, un veneno cuya concentración varía entre las cosechas.

Una firma que procesa granos ha recibido 8 diferentes cosechas, pero la distribución de aflatoxin, en partes por billón, debe ser estimada antes de que sean aceptadas.

Este ejemplo usa el archivo aflatoxin.sav cuyos datos consisten en 16 muestras de cada una de las entregas de las ocho en que se han transado la operación; traemos el archivo

Menú → *Analize* → *Descriptive Statistics* → *Explore* → *Aflatoxin PPB* como la variable dependiente y *Corn Yield* como la variable de factor → OK

Tabla de Resultados 8.4

	Corn Yield			Statistic	Std. Error
Aflatoxin PPB	1	Mean		20,2500	1,07819
		95% Confidence Interval for Mean	Lower Bound	17,9519	
			Upper Bound	22,5481	
		5% Trimmed Mean		20,4444	
		Median		21,5000	
		Variance		18,600	
		Std. Deviation		4,31277	
		Minimum		12,00	
		Maximum		25,00	
		Range		13,00	
		Interquartile Range		8,00	
		Skewness		-,788	,564
		Kurtosis		-,655	1,091

De acuerdo con los términos del ejemplo, el límite establecido acepta como máxima una proporción de 20 unidades de aflatoxin por billón.

En la pantalla de resultados tenemos el cuadro general para las 8 entregas (Yields) y las 16 muestras que se ha elegido de cada una.

Luego nos da la tabla con los estadísticos respectivos para cada una de las 8 entregas; los cuadros son muy extensos para copiarlos, pero podemos analizar uno de ellos

La Tabla de Resultados de la primera entrega tiene una Media de 22.25 unidades por billón, al parecer algo superior a los 20 por billón exigidas por los compradores.

Sin embargo, cuando lleguemos al capítulo sobre la Significancia Estadística podremos determinar si ese indicador está o no dentro del rango de la Media

La Tabla Pivoteada

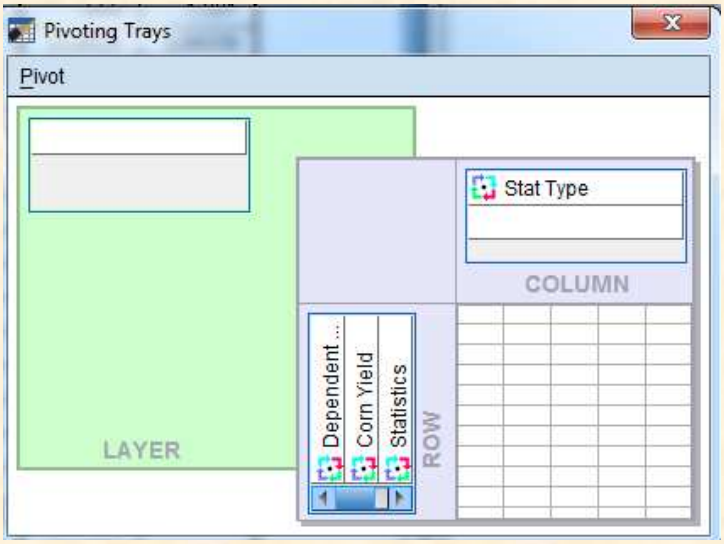
En la Pantalla de Resultados anterior vimos una extensa Tabla de Resultados que incluía los indicadores para las 8 entregas

Nosotros trajimos sólo la primera muestra por lo extenso del cuadro; ahora debemos analizara cuáles entregas tiene menos de 20 unidades por billón

Hacemos doble click en cualquier lugar de la tabla grande, la que se hará un tanto oscura y aparecerá una nueva; en el menú de esa tabla hay la opción Pivote, la activamos.

Hacemos click en Pivote Trays, la activamos y tenemos la siguiente bandeja:

Bandeja 8.1



Con el cursor arrastramos la etiqueta Statistics a la casilla Stat Type
De inmediato aparecerá la tabla pivoteada que es demasiado extensa para traerla

La tabla pivoteada nos muestra, en forma horizontal, todos los indicadores estadísticos que hemos estudiado hasta ahora,

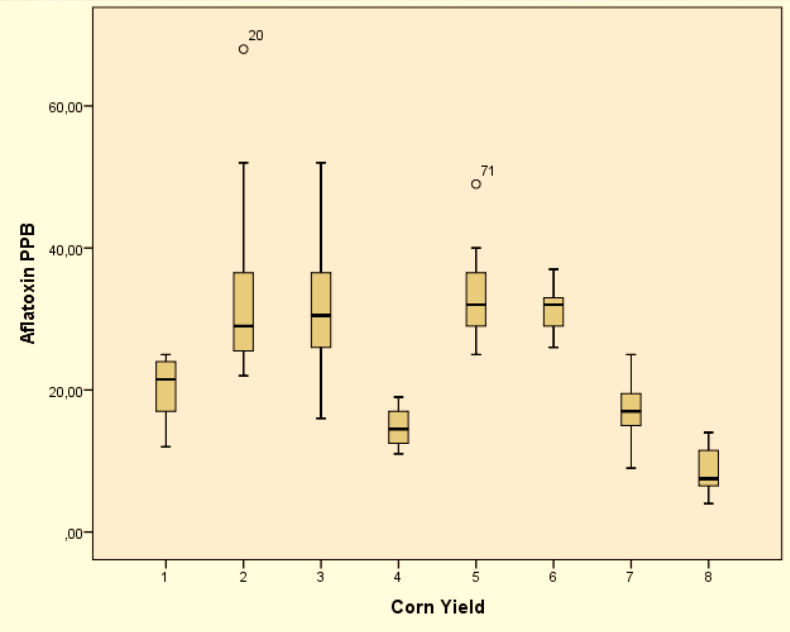
En esa tabla observamos que de acuerdo con la media, sólo las entregas 4, 7 y 8 caen por debajo de las 20 unidades por billón establecidos.

Pues las medias aritméticas de cada una de ellas, son menores a 20 unidades de afloto-xin por billón, que era el requerimiento exigido.

Necesitamos ahora, los Diagramas de Caja para visualizar mejor estos datos

Menu → Descritive Statistics → Explore → Options → Descritiv → Outliers

Diagrama de Caja 8.2



La Pantalla de Resultados muestra dos tipos de información, de las cuales, por el momento, nos interés sólo los *Diagramas de Caja* para cada una de las muestras

Las características de los Diagramas de Caja fueron analizadas en el anterior capítulo
Vemos que hay valores extremos, los que merecen una atención especial.

Ya lo dijimos, la línea gruesa dentro de una caja marca la *Mediana* de la distribución, que es también el percentil 50 de esa misma distribución.

Ejemplo, la Mediana del nivel de *aflatoxin* en el grupo 1 es 21.50 por billón
Observemos que las medianas varían considerablemente entre los diagramas de caja.

Los hinges
Son los límites de las cajas de cada variable.
Los límites inferiores (hinges) de las cajas marcan el percentil 25.

Las líneas superiores (hinges) de las cajas marcan el percentil 75 de cada distribución.

Para el grupo 1, v.g, el valor inferior de la caja es de 17,9519 unidades aflotoxin por billón y el valor superior es 22,5481 unidades de aflotoxin por billón

Para el grupo 3, el valor inferior es 27.1954 y el superior es 38.1796

En la tabla grande, estos valores corresponden a los límites inferiores y superiores de los intervalos de confianza de cada muestra.

Los Whiskers

Son las líneas verticales que parten de los hinges

Los whiskers aparecen encima y debajo de los límites de las cajas, es decir, por encima y debajo de los *hinges;* para el grupo 2 el valor mínimo es 22 y el máximo es 52.

Los valores *outliers,* están identificados con la letra mayúscula "O".
El grupo 2 tiene un valor *outlier* de 68 y está etiquetado con 20

Esto nos dice que el elemento número 20 de la muestra tiene un valor de 68, es decir, un valor por encima de los demás; es un *outlier.*

El grupo 5 tiene un valor *outlier* de 49, etiquetado con 71; nos indica que el elemento número 71 de la muestra tiene un valor 49, que está muy alejado de los demás.

Valores extremos y ourliers

Los valores extremos son identificados con asterisk (*)
No hay valores extremos en estos datos.

Explorando los datos de las Muestras

Ahora conoceremos los que es "capa" (layer) el uso del estadístico "hojas y tallos" y la manera de establecer la cercanía o alejamiento de una distribución a la normal.

Siguiendo nuestra metodología, lo haremos por medio de ejemplos concretos.

Una compañía manufacturera usa nitrido de plata para fabricar soportes de cerámica que deberán resistir temperaturas de 1500 grados centígrados o mayores.

Se sabe que la distribución de una aleación "estándar" de resistencia al calor es normal; pero ahora se prueba una nueva aleación "premiun".

Asumimos que su distribución no es conocida.

Se nos pide comparar los resultados y, en su caso, las diferencias sobre las características entre dos muestras logradas en la fábrica.

Plantearemos un problema específico y una vez que tengamos los datos, los interpretaremos en detalle; para el problema usaremos la opción Explorar.

Traemos el archivo ceramics.save

→*Alloy (Aleación)* como el factor variable →*labrunid* (como la etiqueta de casos)
→statistics: descriptive → Continue → OK

Tabla de Resultados 8.5

	Alloy			Statistic	Std. Error
Degrees Centigrade	Premium	Mean		1542,07	,61165
		95% Confidence Interval for Mean	Lower Bound	1540,87	
			Upper Bound	1543,28	
		5% Trimmed Mean		1541,280	
		Median		1539,711	
		Variance		89,789	
		Std. Deviation		9,47569	
		Minimum		1530,44	
		Maximum		1591,04	
		Range		60,61	
		Interquartile Range		11,51	
		Skewness		1,439	,157
		Kurtosis		3,036	,313
	Standard	Mean		1514,65	,62004
		95% Confidence Interval for Mean	Lower Bound	1513,43	
			Upper Bound	1515,87	
		5% Trimmed Mean		1514,73	
		Median		1514,53	
		Variance		92,269	
		Std. Deviation		9,60566	
		Minimum		1488,30	
		Maximum		1537,99	
		Range		49,69	
		Interquartile Range		13,51	
		Skewness		-,078	,157
		Kurtosis		-,343	,313

→*Menu* → *Analize* → *Descriptive Statistics* → *Explore* →
→*Degrees Centigrade* (como la variable dependiente)

En en la Tabla de Resultados 8.5, la columna Alloy (Aleación) sirve de capa (Layer) a los dos procedimientos: el Standard y Premiun.

La columna de Degrees Centigrade, muestra los estadísticos de las temperaturas que cada tipo de aleación (Estándar y Premiun) alcanza

Por ejemplo, la Media del método Premiun es 1542,07.
Su intervalo de confianza: el límite mínimo es 1540,87 y el límite máximo es 1543,28

Esto quiere decir que el método Premium soporta temperaturas hasta 1543.28 grados

Por su parate, la Media del método Standard es 1514,65
Su intervalo de confianza: el límite mínimo = 1513.43 y el límite máximo es 1515.87

Pero el valor de las Medianas muestra un significado inverso

Para el soporte *Premiun*, 1539,72
Para el soporte *Standard,* 1514,53.

Pero la simetría y la curtosis son menores en el soporte Standard.
Esto nos sugiere que la distribución Standard tiende más a la Normal que la Premiun.

La diferencia entre la Media y la Mediana, en Premiun, parece significativa.
En la segunda parte de esta obra aprenderemos a usar las Pruebas de Normalidad.

Mientras tanto, analicemos los estadísticos que nos muestra la tabla de resultados.

Observemos que la tabla de estadísticos divididos por soporte Premiun y Standard del anterior capítulo incluye la Media Recortada (Trimmed mean)

La Media Recortada deriva de la exclusión del 2.5% de los valores más bajos y del 2.5% de los valores más altos de cada muestra

Eso es lo que significa el 5% escrito al lado.
Lo hace así para omitir los valores extremos que podrían deformar la muestra.

Si el valor de la *Trimmed mean*, excluido el 5% de los valores extremos, es muy diferente de la media, sabremos que nos encontraremos con valores outliers.

La Media recortada en el soporte Premium, 1541,28 es mayor que la de Standard 1514,73 esto es, el soporte *Premiun* tiene mayor resistencia al calor que el *Standard.*

La Pantalla de resultados también desplegó los Diagramas de Caja individuales para ambas aleaciones *Premiun y Standard* con sus respectivos *outliers* y *valores extremos*

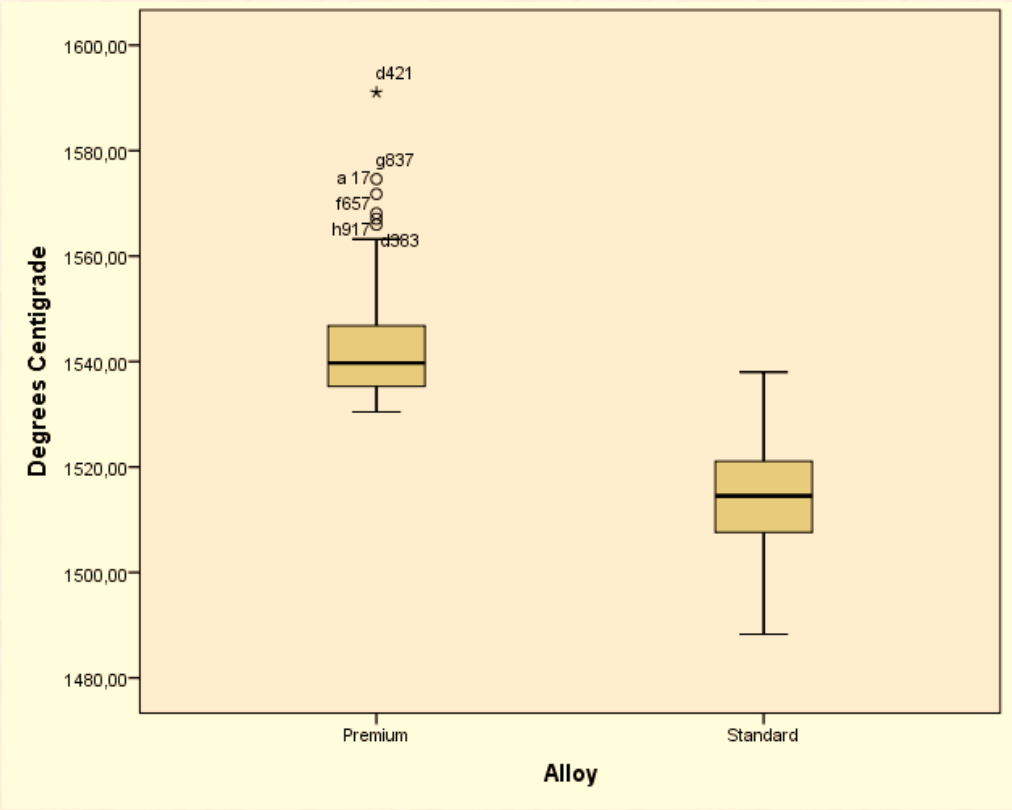
Estos datos apoyan lo que ya sabíamos por los anteriores indicadores:

Los datos de la distribución sobre la resistencia al calor del soporte S*tandard* tienen una distribución más cercana a la normal, que las aleaciones del soporte *Premiun.*

Los diagramas de caja también despliegan valores extremos y outliers mayores y menores, clasificados por Alloy (aleación) para ambos soportes *Premiun y Standard*

Esto se observa en el Diagrama de Caja 8.4

Diagrama de Caja 8.4



Los Diagramas de Caja 8.4 nos sugieren que dada la presencia de outliers en el soporte Premiun es necesario comprobar si ambas muestras tienen la misma varianza.

Ése es un aspecto cuyo estudio corresponde al capítulo *Análisis de la Varianza* ANOVA, el que será incluido con detalle en la segunda parte de esta obra.

Es decir en lo que corresponde a Inferencia Estadística

Hojas y Tallos (Stem-and-Leaf Plot)

Por ahora, vayamos al nuevo Estadístico: Hojas y Tallos (*Stem-and-leaf plots)* opción que nos muestra los valores exactos de las observaciones individuales.

No sólo de las medias o de otro tipo de estadísticos, sino de todos y cada uno de los valores individuales, algo que es de gran interés en los controles de calidad.

Para visualizar mejor sus características, repitamos el ejercicio anterior.

Traigamos el archivo *ceramics.save*

Repitamos el proceso, pero en la opción gráficos del cuadro de diálogo de Explore, pulsemos la casilla Plots → Stem-and-Leaf plots → Continue → OK

En la pantalla de Resultados aparecerá la información requerida.
Empezaremos el análisis del soporte Premium.

Interpretación de la gráfica Tallos y Hojas
La columna izquierda registra las frecuencias de cada nivel de temperatura

En total suman 240 valores, los que coinciden con el número que nos da el cuadro de *Resumen del procesamiento de los casos*

Estadístico Tallo y Hojas
(Stem-and-Leaf Plot for batch= Standard)

Frequency	Stem & Leaf	
2,00	148.	88
2,00	149.	44
12,00	149.	566677788999
22,00	150.	0001111122333333333444
35,00	150.	55555555666667777777777777888888999
54,00	151.	000000001111111111112222222333333333333333444444444444
43,00	151.	55555566666666677777777777777778888899999999
32,00	152.	00000000011111122223333333334444
22,00	152.	5555555666667777888899
15,00	153.	000000122223344
1,00	153.	7

Los datos de columna del medio, se llama S*tem* y los de la última columna, Leaf.
Escojamos, al azar, una fila, digamos la que muestra una frecuencia de 22.00

La columna de Stem registra 150 y la columna de *Leaf* los valores que completan el valor 150 de la columna Stem, empezando por 0.

Escojamos el primer "0" de *Leaf* y lo añadamos a 150 de *Stem:* tendremos 1500; hay 3 ceros en la columna *Leaf*, habrá 3 casos que soportan temperaturas de 1500 grados

Escojamos ahora el valor de Stem de 151; hay dos frecuencias para ese valor, que son 54 y 43, datos a la izquierda de 151 y de 151; elijamos el que tiene la frecuencia de 54.

Agregamos el primer 0 de la columna de *Leaf* al valor de 151, éste se convertirá en 1510; como hay 8 ceros en la columna *Leaf*, entonces habrá 8 observaciones de 1510

Es decir, 8 casos que registran una capacidad de soportar 1510 grados centígrados.

Ahora, en vez de completar el valor *Stem* de 151 con cada cero, haremos lo mismo, añadiendo al valor Stem el primer 1 de la columna *Leaf.*

El valor Stem se convertirá en 1511; como hay doce valores "1", entonces sabremos que habrá doce observaciones que representan soportes de 1511 grados.

Hacemos lo mismo con los valores "2", "3" y "4", respectivamente.

El resultado nos dará 7 valores de 1512, 15 valores de 1513 y 12 valores de 1514 que soportan esas temperaturas medidas en grados centígrados.

El valor de la primera columna a la izquierda, Frecuency, nos dice que en total hay 54 soportes que resisten temperaturas desde 1510 hasta 1514 grados centígrados.

Esto quiere decir que la mayor parte de los elementos de Standard soportan temperaturas entre 1505 y 1519 grados.

Los datos nos indican que los valores se concentran, entre 1505 y 1519 grados.

Estadístico de Tallo y Hojas
(Stem-and-Leaf Plot for batch= Premium)

Frequency	Stem &	Leaf
24,00	153.	000000011111111111111111
22,00	153.	2222222222333333333333
26,00	153.	44444444445555555555555555
26,00	153.	66666666666667777777777777
24,00	153.	888888888888999999999999
19,00	154.	0000000000111111111
25,00	154.	22222222222222333333333333
10,00	154.	4444455555
12,00	154.	666666667777
10,00	154.	8888999999
8,00	155.	00111111
4,00	155.	2223
6,00	155.	445555
6,00	155.	666667
6,00	155.	888899
3,00	156.	011
3,00	156.	223
6,00	Extremes	(>=1566)

Realizaremos el mismo análisis de los datos con la tabla Stem-and-Leaf para los soportes *Premium.*

El gráfico para Premium nos muestra, por ejemplo, que hay 24 observaciones que soportan desde 1538 hasta 1539 grados centígrados de calor.

Habrá 8 muestras que soportan desde 1550 hasta 1551 grados centígrados de calor.

Procedimientos relacionados

Podemos usar el procedimiento Frequencies para resumir y estandarizar variables de escala y variables categóricas

También podemos usar el procedimiento Means que provee estadísticas descriptivas y ANOVA para estudiar la relación entre variables de escala y categóricas.

El procedimiento Summarize provee estadísticas descriptivas y resúmenes de casos para examinar las relaciones entre variables de escala y categóricas.

El procedimiento OLAP Cubres nos proporciona estadísticas descritivas para analizar las relaciones entre variables escala y categóricas.

Por último, el procedimiento Correlations provee resúmenes que describen la relación entre dos variables de escala.

Todos esos procedimientos serán analizados en la segunda parte de este Manual.

Aleluya: con esto hemos terminado la primera parte de este curso, es decir, hemos concluido el programa de la Estadística Descriptiva y la Probabilidad.

La Estadística Descriptiva es el pivote central en el que descansan las otras ramas de nuestra disciplina; espero que la hayan asimilado muy bien, pues nos será muy útil.

Con estos conocimientos adquiridos estamos listos para empezar con la segunda parte del Manual, esto es, **Inferencia Estadística.**

2. INFERENCIA

9
PRUEBAS DE HIPÓTESIS

La prueba de hipótesis forma parte de la Inferencia Estadística.
Se define la Inferencia Estadística como:

> *Una de las dos ramas fundamentales de la Estadística cuyo objeto es llegar a conclusiones sobre las características de una población, usando como escenario de análisis la muestra que se ha obtenido de esa población*.

El Teorema del Límite Central, como vimos en la parte primera, establece que una muestra adecuada refleja las características de la población respectiva.

La Inferencia Estadística tiene un conjunto ya probado de técnicas para medir el grado de certidumbre de las conclusiones a partir del estudio de la muestra.

Queda claro, a partir de la definición, que si en vez de la muestra se tiene los datos completos de la población, la Inferencia Estadística deja de tener sentido.

Por otro lado, a diferencia de las matemáticas, la estadística no es una ciencia exacta.
Al contrario, siempre toma en cuenta la probabilidad de error.

La Inferencia Estadística acepta dos divisiones principales:

La Inferencia Paramétrica

Opera sobre muestras cuyas poblaciones tienen una forma determinada de distribución (La Normal, la de Poisson, la Binomial…)

En este tipo de inferencia se conoce las distribuciones de los valores pero se ignoran los parámetros que son, precisamente, los que deseamos saber

Inferir sobre los valores de los parámetros de la población, a través del análisis de una muestra, es la tarea principal de la Inferencia Estadística.

La Estimación de los parámetros puede ser puntual o por intervalos.

Por otra parte es necesario tener presente que hay varias clases de muestra para el análisis y las inferencias respectivas acerca de la población.

De todas ellas, en este curso usaremos la muestra aleatoria, tal como quedó definida en capítulos anteriores.

Muestras pequeñas

Por lo general, las muestras que se obtienen en los trabajos de campo son pequeñas y no se conoce la desviación estándar de la población.

Por este motivo es que nuestro análisis se concentrará en las muestras pequeñas, aunque no hay ninguna diferencia de procedimiento con las muestras grandes.

Cuando nos referimos a las muestras pequeñas, no vamos a la curva normal sino a la curva de distribución "t".

La Muestra Aleatoria

La teoría de las muestras establece que de una población dada es posible extraer un número muy grande de muestras del mismo tamaño.

Si todas y cada una de esas muestras tienen la misma probabilidad de ser escogidas, entonces tendremos una Muestra Aleatoria.

Del mismo modo, todos los elementos que componen una muestra aleatoria también son aleatorios, es decir, todos tienen la misma probabilidad de ser escogido.

Con estas breves consideraciones ya podemos encarar el nuevo tema.
Para empezar nuestro trabajo es preciso definir algunos términos.

Hipótesis Nula

Es la suposición que deseamos testar; se simboliza con Ho.

Supongamos que deseamos testar la hipótesis de que la media aritmética del ingreso de una de una población es de $500 semanales.

Escogemos una muestra aleatoria de esa población.

La manera de representar esta suposición es: Ho: $\mu = 500$ = Hipótesis Nula
Supongamos que hay algunos indicios para dudar de que la media sea 500.

Para asegurarnos, contrastamos la Ho con otra, llamada Hipótesis Alternativa

Hipótesis alternativa

Se simboliza como H_1; esta hipótesis puede presentarse de tres maneras:

H_1: $\mu > 500$; la media, según la hipótesis alternativa, es mayor que 500.
H_1: $\mu < 500$; la media según la hipótesis alternativa es menor que 500.
H_1: $\mu \neq 500$; la hipótesis alternativa dice que la media no es igual a 500

Una vez que tenemos planteadas las hipótesis es preciso escoger un criterio que nos permita aceptar o no aceptar la hipótesis nula con un nivel adecuado de confianza.

Origen del Test de Hipótesis

Los primeros intentos sobre el test de hipótesis vienen del año 1279 en Inglaterra.

Cada vez que se forjaba monedas, un número pequeño de ellas eran depositadas en una caja cerrada; un jurado de herreros tomaba una de las monedas al azar.

Luego comparaba si la moneda entraba en una zona de tolerancia con relación a su peso y la ley del oro que por ley debía contener.

En cada uno de los intentos, se suponía dos hipótesis

Ho: Las monedas están en la zona de tolerancia
H_1: Las monedas están fuera de la zona de tolerancia

Son las que ahora conocemos como nula y alternativa, respectivamente.

El Nivel de Significancia

Recordemos que lo que estamos testando es la diferencia de dos medias: la que se supone verdadera, planteada por la Ho y la que obtenemos de la muestra.

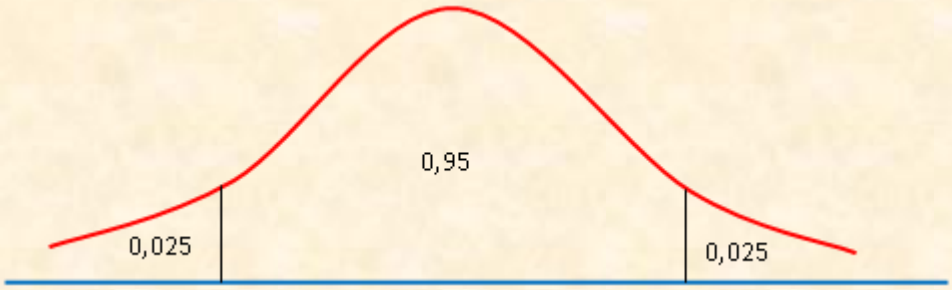
Pero es preciso aceptar que habrá un margen de error en la estimación.

Ese margen de error determina si se acepta o no que la diferencia entre las dos medias sea estadísticamente significativa; supongamos que aceptamos un error del 5%

Esto quiere decir que aceptamos la probabilidad teórica de que de cada cien muestras, cinco estén fuera de los límites establecidos al plantear la hipótesis.

Ahora bien, en este caso, si la diferencia de las medias es estadísticamente significativa, es decir, mayor del 5%, rechazaremos la Ho.

Gráfica 9.1
Curva distribución t



Con el objeto de visualizar la interpretación del nivel de significancia, en este caso, del 5%, se diseña la Gráfica 9.1, que es una curva de distribución "t"

Asumiremos que la Hipótesis alternativa: $H_1 \neq \mu$, es decir, una alternativa de dos colas, por lo que la Gráfica 9.1, se divide en tres partes.

La primera, la más extensa, es el área de aceptación, esto quiere decir que si el estadístico estimado cae en esa área, aceptaremos la Ho. Por otro lado, cada una de las colas corresponde a la mitad del error del 5%, por lo tanto, en cada una habrá el 2.5% de cometer el error a partir de la vertical.

Si el valor del estadístico que vamos a estimar cae entre una de las dos colas, rechazaremos la Ho, pues estará fuera del área de aceptación. Si escogiéramos un nivel de significancia del 10% de dos colas, entonces el área de aceptación sería 90% y cada cola tendría el 5% de probabilidad de error.

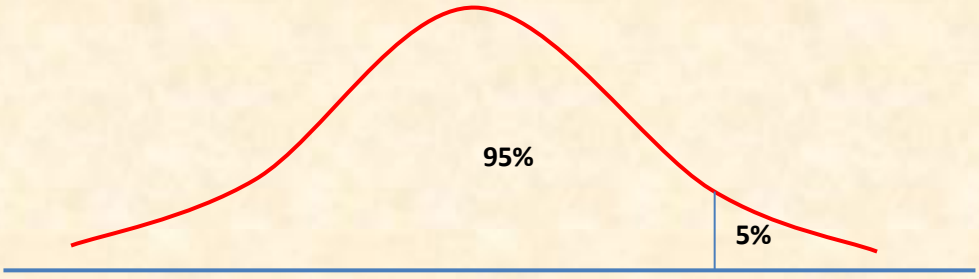
Por otra parte, si la H_1 postula que la media es mayor que la planteada por la Ho entonces tendremos un test de una sola cola a la derecha, como la *Gráfica 1.2.*

Si la H_1 planteara que la media es menor que la postulada por la Ho, el 5% de error se concentraría en la cola izquierda y el área de aceptación continuaría 95%

Si el nivel de significancia es del 5%, el área de aceptación será del 95% y habrá una sola cola, la que estará a la derecha de la curva, conteniendo el 5% del error

Si el estadístico que vamos a estimar cae en la cola del 5%, rechazaremos la Ho dado que ese valor está en la zona de error.

Gráfica 1.2
Una sola cola



En Síntesis
Para llevar a cabo un test de hipótesis procedemos del siguiente modo

Primero, establecer una hipótesis nula: Ho, la que será testada por la H_1
Segundo, establecer un nivel de significancia o error (el 5% o el 1%)

Para realizar ejercicios manuales antes de ingresar al SPSS, traeremos la Tabla "t"

Esta Tabla es la que se usa para muestras.

Tabla 9.1
La distribución "t"

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	**1.729**	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646

40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Uso de la Tabla "t"

Ahora aprenderemos a usar los valores de la tabla "t", por medio de ejercicios que nos permitan familiarizarnos con este importante estadístico.

Procederemos del mismo modo que procedimos con la Curva Normal; es decir, antes de recurrir al SPSS, aprenderemos a usar manualmente la tabla "t".

El principio es similar al de la tabla "z" o normal que vimos en capítulo anterior.

En la parte superior se muestra dos filas: para una cola y para dos colas.
En cada caso, los encabezamientos de las columnas indican el nivel de confianza.

Si el número de elementos de una muestra es > 30, la distribución tiende a la Normal, por lo que podríamos usar la tabla para la normal en vez de la presente "t".

Ejercicio 1

Una maestra de escuela dice que el promedio de calificaciones de sus alumnos es de 90 puntos; el director desea saber si está en lo cierto.

El Director nos pide que realicemos los estudios necesarios para confirmar o no lo que ha informado la maestra.

Nos pide un margen de error del 10%, es decir, un nivel de significancia del 10%.
De inmediato deducimos que el nivel de confianza es del 90%.

Empezamos obteniendo una muestra, digamos, de 20 alumnos del total.

Revisados los valores encontramos que el promedio de notas, es 84, que es menor que 90, aunque no sabemos si es significativamente menor.

No basta que un valor sea diferente del que plantea la Ho.

Por ejemplo, si hubiéramos obtenido una Media muestral de 89.98: ¿Diríamos que es diferente del nivel, 90, que deseamos testar?

Es necesario contar con un referente objetivo.

Un referente que nos indique si hay realmente una diferencia, *estadísticamente* significativa, entre ambos valores

Luego veremos que el nivel Sig. del SPSS es el que nos da el referente que buscamos para esta clase de problemas, pues es objetivo, no depende del juicio de nadie.

Pero sigamos con nuestro ejercicio manual; supongamos que, usando los datos de la muestra, hemos estimado que su desviación estándar es: $\sigma = 11$

Antes de ir a la tabla 9.2, ordenamos los datos

$\mu = 90$ la media de la población, sustentada por la profesora.
$X^* = 84$, que es la media de la muestra de los 20 elementos
$\sigma = 11$
n = el número de elementos en la muestra

El director nos pidió que hiciéramos un test de hipótesis con un 10% de error máximo, es decir, con un nivel de significancia del 10%.

El nivel de significancia del 10% establece el nivel de confianza del 90%.
Este nivel de confianza es el que aparece en las dos primeras filas de la tabla "t".

α = Nivel de Significancia: 10%.
90% Nivel de Confianza

Dado que el director no nos dijo si él creía que la nota fuera mayor o menor a 90, asumimos que el planteamiento correcto en este caso es:

Ho: $\mu = 90$
H_1: $\mu \neq 90$

Es decir, decidimos que vamos a testar un test dos colas, puesto que H_1, la hipótesis alternativa, plantea que la media puede ser mayor o menor que la hipótesis nula, Ho.

Busquemos el nivel de significancia del 10%, entonces la columna que registra el 90% será la apropiada, pues es el nivel de confianza pertinente.

Dado que el problema que vamos a resolver es de dos colas, al 10% de significancia, entonces escogeremos la segunda fila del encabezamiento de la tabla.

Los valores a lo largo de cada una de esas filas son los niveles de confianza para cada caso: una cola o dos colas (One Sided or Two Sided)

Nosotros usaremos los estadísticos para *dos colas*, pues eso es lo que dice H_1

En la primera columna de la tabla "t" se registra el número de elementos de las muestras, hasta 30, que es el máximo número que se opera con la tabla "t"

En las filas de cada número de elementos, están los valores "t" en cada caso.
El número de elementos en la muestra que tenemos es 20.

Pero la fórmula con que fue construida la tabla "t" disminuye 1 elemento a cada muestra, por lo tanto debemos operar con un valor de 19.

Identificamos el valor 19 en la primera columna.

Recorremos la fila donde está el valor de 19 hasta encontrar el valor bajo la columna 90%, que corresponde a "dos colas".

Recordemos que el director nos pidió una prueba de hipótesis con un nivel de error máximo (nivel de significancia) del 10%;

Ese valor corresponde a dos colas, porque el de una cola para 19 elementos, está en la misma columna pero en la primera fila con un nivel de confianza de 95%.

Cada vez que tengamos problemas que plantean una hipótesis alternativa de dos colas, nos fijamos en la segunda fila del encabezamiento de la tabla.

Recorremos la fila hasta ponernos en la columna 90%, en este caso, cuyo valor coincide con el de 90% de la segunda fila de la parte superior de la tabla.

Ese 90% es el nivel de confianza, emergente del nivel de significancia del 10% que estamos usando; el valor es 1.729, que recibe el nombre de *valor crítico.*

Ya hemos logrado la primera parte de la tarea.

La segunda parte de nuestra tarea es saber si el valor "t" que vamos a estimar está por encima o por debajo de 1.729, que hemos identificado en la tabla.

Comprobación manual

Si el resultado que vamos a obtener manualmente es menor que ese valor crítico, aceptaremos la hipótesis nula, el promedio de las evaluaciones es, del 90%.

Si el resultado es mayor, rechazaremos la hipótesis nula.
(Cuando el resultado pasa de 3 la rechazamos automáticamente)

Ahora bien, para testar si la diferencia de medias aritméticas de la población y de la muestra difieren significativamente o no, usamos la fórmula "t"

$$t = (X^* - \mu)/\sigma_x$$

Es la versión para pequeñas muestras de la fórmula "z" que habíamos usado en la parte Primera del curso, para calcular los resultados, antes de usar el SPSS.

Esta fórmula nos permite contrastarla la media que estimamos en la muestra (X* = 84) con la Media que nos da la Ho (μ = 90)

Para eso es preciso estimar antes el estadístico σ_x

En la igualdad 1.2 vimos que: $\sigma_x = \sigma/\sqrt{n}$
Supusimos que habíamos estimado la desviación estándar de la muestra: $\sigma = 11$

El error estándar de la Media será; $\sigma_x = 11/\sqrt{20} = 11/4.47 = 2{,}46$
La "t" estimada, usando la fórmula, es $t = (84 - 90)/2.46 = -6/2.46 = -2{,}44$

El signo es negativo porque que el promedio de la muestra, = 84, es menor que 90
Pero el valor absoluto es mayor que el valor de la tabla.

Esto es, 2.44 > 1.729 en valores absolutos

En efecto, el nuevo valor absoluto de "t", -2.44 es mayor que el valor de la tabla, el cual era de 1.729, por lo que rechazamos la hipótesis nula de la Directora

Conclusión
La media aritmética de las calificaciones (de la población) es menor a 90

De esta manera, resolvimos manualmente un problema-tipo de prueba de hipótesis para una sola media cuando las muestras son pequeñas.

Es decir, cuando tienen menos de 30 observaciones.
Complementamos estos conceptos con los siguientes.

Error de Tipo I
Lo cometemos cuando la hipótesis nula es cierta y sin embargo la rechazamos.

Error de Tipo II
Se comete cuando se acepta una hipótesis nula siendo que ésta es falsa.

Poder de una Prueba de Hipótesis
Sobre la base de los tipos de error, se define el Poder de una Prueba de Hipótesis a la probabilidad de rechazar la hipótesis nula cuando es realmente falsa.

Se refiere a la capacidad de un test para detectar adecuadamente las diferencias de dos medias, la propuesta por la Ho y la H_1 como estadísticamente significativas.

Símbolos utilizados
La media aritmética de la *Población* se representa por la letra griega: μ.

Para le Media Aritmética de la *Muestra*, usaremos X*

Con estos símbolos, podemos establecer que la Hipótesis Alternativa puede tener tres significados diferentes, de acuerdo con el criterio del investigador.

H_1: X* ≠ μ; la media muestral (X*) es diferente de la media de la población (μ)
$H_{1:}$ X* > μ; la media muestral es mayor que la media de la población.
H_1: X* < μ; la media muestral es menor que la media de la población

También debemos tomar en cuenta un concepto muy importante:
σ = Desviación estándar de la muestra

Es la suma del cuadrado de las distancias entre cada observación de la muestra y su media, dividido entre el número de elementos de la muestra menos 1.

$$\sigma = \sqrt{\Sigma(X_i - X^*)^2/(n-1)} \qquad (9.1)$$

En esta fórmula, el símbolo X_i es el valor de cada elemento de la muestra; X* es la media aritmética de la muestra y n es el número de elementos en la muestra.

El símbolo Σ indica la suma de todas las diferencias dentro del paréntesis, las que previamente han sido elevadas el cuadrado.

Otro concepto que debe captarse muy bien es el siguiente:

$$\sigma_x = \text{Error estándar de la Media} = \sigma/\sqrt{n} \qquad (1.2)$$

Es necesario notar que σ_x, error estándar de la Media, en (1.2) es diferente a σ, que es la desviación estándar de la muestra, tal como queda establecido en (1.1)

σ_x = Error estándar de la Media
Ahora vamos a definir este parámetro.

Si tenemos una población de 5000 elementos y deseamos lograr una muestra de 50, escogeremos los cincuenta elementos de manera aleatoria.

Luego, volvemos a escoger otra muestra, también de 50 elementos.

Con toda seguridad que los valores de los elementos de la segunda muestra serán diferentes de los valores de la primera muestra, aunque el número de elementos sigue 50.

Así, podemos sacar muchas muestras de 50 elementos, las que diferirán una de las otras; sería una coincidencia que sacáramos muestras con los mismos elementos

Si tomamos el número teórico de muestras de 50 elementos de una población de 5000 habrá una Media Aritmética del total de las muestras.

Por eso es que habrá también una desviación estándar de las distribuciones muestrales

Esa desviación estándar de las muestras con relación a la Media del conjunto de todas ellas es que se llama: σ_x = Error estándar de la Media

La σ nos muestra las diferencias de cada elemento de la muestra con relación a la Media Aritmética de esa Muestra, en particular.

En cambio, σ_x es la desviación estándar de las medias de todas las muestras posibles de igual tamaño que, teóricamente, se puede extraer de una población.

Con el SPSS

El primer procedimiento del SPSS para testar las hipótesis con la distribución "t" se denomina procedimiento de una sola muestra.

Un fabricante de automóviles de alto rendimiento produce frenos de disco que deben medir 322 milímetros de diámetro.

Los analistas toman muestras al azar de 16 discos realizados por cada una de ocho máquinas de producción, para medir sus diámetros.

Vamos a utilizar el Test "T" para determinar si los diámetros medios de los frenos en cada muestra difieren significativamente de los 322 milímetros requeridos.

Una variable nominal, Número de máquina, identifica la máquina de producción usado para hacer el freno de disco.

Los datos de cada máquina deben ser probados como una muestra separada, por ello vamos a dividir el archivo numerando cada una las máquinas que dieron las muestras

El Procedimiento T test

Con este procedimiento vamos a establecer la diferencia entre la Media en cada muestra y un valor previamente hipotetizado, en este cso, 322 mm

Traemos a la pantalla del SPSS el archivo *brakes.save*

La variable nominal, Número de Máquina, identifica la máquina que ha producido los discos de freno, en cada una de las 8 máquinas productoras

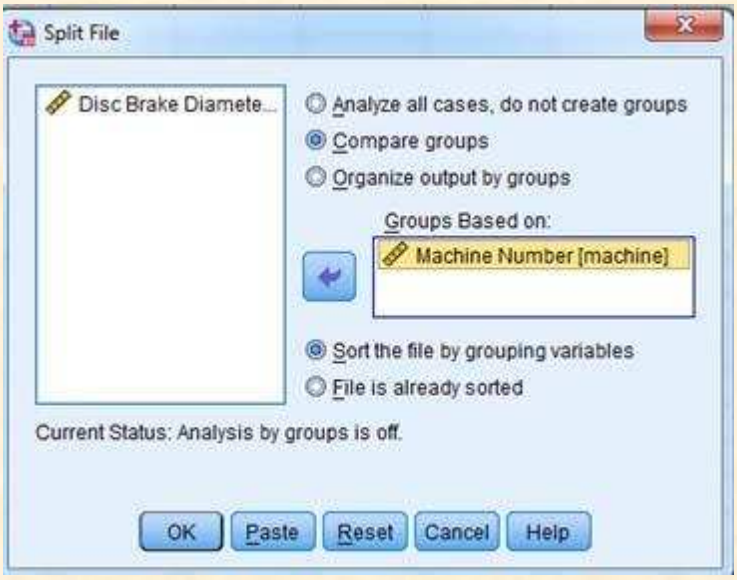
Debido a que los datos de cada máquina deben ser testados como muestras separadas, el archivo debe ser dividido en grupos, identificando cada máquina con un número

Para divider el archive:

Menu → Data → Split File → Compare Means

Con la fleche de dirección introducimos en la pantalla la variable Machine Number
La Pantalla de Resultados del SPSS trae la Caja 9.1

Caja 9.1



Ahora vamos a comparar las medias de cada máquina con el valor de hipótesis 322 mm

Menu → Analyze → Compare Means → One-Sample T Test→ Disc Braque Diameter → 322 en la casilla de Tested Value → Options → Continue → OK.

Tabla de Resultados 9.1

Machine Number		N	Mean	Std. Deviation	Std. Error Mean
1	Disc Brake Diameter (mm)	16	321,998514	,0111568	,0027892
2	Disc Brake Diameter (mm)	16	322,014263	,0106913	,0026728
3	Disc Brake Diameter (mm)	16	321,998283	,0104812	,0026203
4	Disc Brake Diameter (mm)	16	321,995435	,0069883	,0017471
5	Disc Brake Diameter (mm)	16	322,004249	,0092022	,0023005
6	Disc Brake Diameter (mm)	16	322,002452	,0086440	,0021610
7	Disc Brake Diameter (mm)	16	322,006181	,0093303	,0023326
8	Disc Brake Diameter (mm)	16	321,996699	,0077085	,0019271

Allí observamos las Medias del diámetro de los discos de frenos producidos por las 8 máquinas, de cada una de las cuales hemos traído muestras de 16 discos

Las diferentes Medias parecen diferir poco del valor de 322 mm requeridos
Hay dos indicadores que es preciso analizar

Std. Deviation

Es el promedio de las desviaciones de la Media Aritmética de cada elemento dentro de la muestra a la que nos referimos

Por ejemplo, Std. Deviation para la máquina 1 es 0.0111568, eso quiere decir que los elementos de la máquina 1 difieren de la Media en un promedio de 0.011568 mm

Std. Error Mean

Es la desviación promedio que hay entre el diámetro de los discos producidos en cada máquina, con la Media de las 8 muestras en total.

El valor Std.Error Mean de la primera máquina nos dice que en promedio, la Media Aritmética de los discos que produce es 0.002782 mm distante de la Media Total

De ahí el cambio de nombre, de desviación estándar a *error estándar* en la notación de los indicadores

Una vez que hemos dividido el archivo en 8 máquinas diferentes, de cada una de las cuales se ha tomado una muestra de 16 elementos vamos a testar los diámetros.

Menu → Analyze → Compare Means → One-Sample T Test

Introducimos Disc Braque Diameter en la pantalla de la derecha →Options → 90% Confidence Intervale → Exclude cases analysis by analysis → Continue → OK

Analizaremos la Tabla de resultados, columna por columna, empezando por la diferencia de Medias (Mean Difference) de la producción de cada máquina con los 322 mm

Mean Difference

Las máquinas 1, 3, 4 y 8 tienen una diferencia de la Media con el signo negativo
Esto quiere decir que, en promedio, la media es un tanto menor a los 322 mm

Lo contrario sucede con las máquinas en los que las diferencias de la Media tienen signo positivo con relación a los 322 mm.

La Columna t

Al poner como valor de referencia 322 mm para el diámetro exigido por las condiciones del problema, implícitamente nos plantea las siguientes hipótesis

$Ho = 322$
$H_1 \neq 322$

El valor "t", cuyo significado analizamos en el anterior capítulo, nos nos da una pauta sobre los valores que estarían por encima o por debajo de Ho.

Sig 2 Tailed

Es la que nos hace saber si el valor "t" en cada caso es o no estadísticamente significativa de la Ho; en este caso, con un nivel de significancia del 10%

En el caso de la primera máquina, el valor "t"es -0.533
A primera vista podemos ver que ese valor es muy bajo

Para constatarlo, vemos que su valor Sig. 2 tailed es de 0.602
Esto quiere decir que =.602 es mayor que el 10% requerido

Por lo tanto, rechazamos la hipótesis nula de que no hay diferencia significativa entre el valor 322 y la media de la máquina 1.

Tabla de Resultados

One-Sample Test

Machine Number		Test Value = 322					
		t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
						Lower	Upper
1	Disc Brake Diameter (mm)	-,533	15	,602	-,0015	-,0074	,0044
2	Disc Brake Diameter (mm)	5,336	15	,000	,0145	,0085	,0199
3	Disc Brake Diameter (mm)	-,655	15	,522	-,0017	-,0073	,0038
4	Disc Brake Diameter (mm)	-2,61	15	,020	-,0046	-,0089	-,0008
5	Disc Brake Diameter (mm)	1,847	15	,085	,0042	-,0006	,0091
6	Disc Brake Diameter (mm)	1,134	15	,274	,0024	-,0021	,0070
7	Disc Brake Diameter (mm)	2,650	15	,018	,0062	,0012	,0111
8	Disc Brake Diameter (mm)	-1,713	15	,107	-,0033	-,0074	,0008

En cambio, la máquina 2 nos da un valor "t" de 5.336 y un Sig. = 0.000
Esto quiere decir que el valor Sig es menor que el 10% hemos determinado.

Por lo tanto, rechazamos la Ho y aceptamos la H1 en sentido de que la máquina 2 está produciendo discos de freno cuyo diámetro es mayor que 322 mm.

Los Intervalos de Confianza

Los Intervalos de Confianza al 90% proveen los límites entre los que se encuentra la verdadera diferencia en el 90% de todas las muestras.

Los Intervalos de Confianza tienen un valor de 0.0, cuando hay perfecta coincidencia entre la media propuesta (322 mm) y las que hay en cada muestra.

Si los dos límites son positivos, deduciremos que las máquinas de donde provienen las muestras, están produciendo frenos cuyo diámetro está por encima de 322 mm.

Eso es lo que sucede con las máquinas 2, 5 y 7
Por lo tanto rechazamos la Hipótesis nula en cada una de esas tres máquinas.

Lo contrario sucede con la máquina 4, pues ambos límites de su Intervalo de Confianza para la diferencia de medias son negativos.

Esto quiere decir, que hay evidencia de que la máquina 4 está produciendo frenos cuyos diámetros son menores que 322 mm; también rechazamos la Hipótesis nula

En las restantes cuatro máquinas no hay razones para rechazar la Hipótesis nula.
Por ello, aceptamos que producen diámetros consecuentes con los 322 mm.

Observamos, de paso, que los valores Sig para las máquinas 2, 4, 5 y 7 son menores al 10% del nivel de significancia que se propuso.

En las máquinas para las que aceptamos la Hipótesis nula, los valores Sig son mayores que el 10% propuesto de nivel de significancia.

El *t−test* de una muestra puede ser usado cada vez que las medias de la muestra deban ser comparadas con algún valor propuesto, dado por la Ho.

En todo caso, se asume que la distribución de la población de la cual se deriva la muestra es razonablemente cercana a la normal.

Se debe tomar nota de la asimetría y de los valores extremos.
En el próximo capítulo nos referiremos a otros tipos de "t test".

10
"T TEST" PAREADOS

Concepto

Uno de los diseños más conocidos es el denominado "pre-post", el que consiste en comparar dos medidas del mismo sujeto antes y después de un tratamiento o estímulo

Si el tratamiento no tiene los efectos previstos, la diferencia entre las medias de la muestra antes y después del tratamiento será próxima a 0; se aceptará la hipótesis nula

Por el otro lado, si el tratamiento es efectivo la diferencia de las medias diferirá significativamente de 0; en ese caso, rechazaremos la hipótesis nula.

El T Test para muestras pareadas se usa para testar la hipótesis esta clase de hipótesis

Es decir, testar si hay o no diferencia significativa entre dos variables sobre datos emergentes de dos medidas tomadas sobre el mismo grupo antes y después.

Ejercicio

Un médico evalúa una nueva dieta para sus pacientes con un historial familir sobre enfermedades del corazón.

Para testar la eficacia de la dieta, tiene la participación de 16 pacientes que seguirán la dieta por seis meses.

Los pesos y los triglicéridos son medidos antes y después de la dieta
El médico desea conocer si esas medidas han cambiado luego de la dieta cumplida.

Uso del SPSS

Los datos respectivos están en el archivo *dietstudy.sav*

Usaremos el Test T para muestras pareadas y así determinar si hay diferencias significativas entre los pesos y los triglicéridos antes y después de la dieta.

Analyze → *Compare Means* → *Paired-Samples T Test* → *Triglyceride y Final Triglyceride* como el primer par de variables

→*Weight y Final Weight* como el Segundo par → OK.

La Pantalla de Resultados registra tres tablas
Tomaremos la primera de ellas y luego, las demás

Tabla de Resultados 10.1

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Triglyceride	138,44	16	29,040	7,260
	Final triglyceride	124,38	16	29,412	7,353
Pair 2	Weight	198,38	16	33,472	8,368
	Final weight	190,31	16	33,508	8,377

Los promedios de los niveles de triglicéridos antes y después de la dieta, muestran una diferencia de entre 14 y 15 puntos durante los seis meses que duró la dieta.

Por otra parte, los promedios de los pesos, antes y después de la dieta, tienen una diferencia desde 198.38 a 190.31 libras, respectivamente.

Las desviaciones estándar para las medias antes y después de la dieta muestran que hubo mayor variabilidad entre la pérdida de peso que entre los niveles de triglicéridos

La Correlación de Pearson
Es la que aparece en la Tabla de Resultados

Tabla de Resultados 10.2

Paired Samples Correlations				
		N	Correlation	Sig.
Pair 1	Triglyceride & Final triglyceride	16	-,286	,283
Pair 2	Weight & Final weight	16	,996	,000

La Correlación de Pearson determina el grado de asociación que hay entre dos variables; es simétrica, pues el grado de correlación entre X y Y es igual al de Y y X

Por ello, no pretende establecer causalidad entre las variables que toma.
En la tabla 10.2 la Correlación de Pearson es –0.286 y su nivel de Sig = 0.283 > 0.10

Sobre esos valores podemos decir, con un nivel de significancia del 10%, que la diferencia entre los triglicéridos antes y después de la dieta no es significativa

Sin embargo, a niveles individuales, los datos son inconsistentes, pues algunos rebajaron sus niveles y otros los mantuvieron.

Por el otro lado, la Correlación de Pearson entre las medias del peso antes y después es 0.996, que se acerca a la correlación perfecta.

Eso se comprueba con su valor Sig = 0.000.

De esos valores deducimos que no tenemos razones para aceptar la Ho y afirmar que sí, existe una diferencia significativa entre los pesos antes y después de la dieta.

En la siguiente tabla la columna Mean despliega las diferencias entre las medias vigentes entre triglicéridos y peso antes y después de la dieta.

La columna 3, desviación la estándar para los promedios respectivos.

La columna (4) Std. Error Mean, proporciona la variabilidad esperada de las repetidas muestras tomadas con 16 elementos c/u y la Media total de todas las medias muestrales

Tabla de Resultados 10.3

Paired Samples Test

		Paired Differences					t	df	Sig. 2-tailed
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Triglyceride – Final triglyceride	14,06	46,875	11,72	-10,91	39,04	1,20	15	,249
Pair 2	Weight – Final weight	8,06	2,886	,72	6,525	9,60	11,17	15	,000

Lo dijimos antes: una población permite extraer muchas muestras similares, para las que se espera un error estándar, que es la que se muestra en la columna 4.

El Intervalo de Confianza del 95% de la diferencia provee los límites estimados entre los que se encuentra la diferencia verdadera en el 95% de las muestras aleatorias

El Estadístico "t"

Se obtiene dividiendo la diferencia de las medias entre sus errores estándar
Ejemplo, en el caso de los triglicéridos, dividimos 14.06/11.71 = 1.20.

El Sig = 0.249 para los triglicéridos, mayor que 0.05, indica que no hay una diferencia significativa entre los promedios antes y después de la dieta.

Esto es, con relación a los triglicéridos, la dieta no tiene efecto alguno.

Pero, el valor Sig =0.000 para el peso sí muestra que hay una diferencia significativa entre los promedios del peso antes y después de la dieta.

La dieta es eficaz en la reducción del peso, pero no en la reducción de triglicéridos.

Resumen

El test "t" para muestras pareadas es apropiadO para relacionar 2 promedios muestrales que deben ser comparados entre sí

Pero, se asume que los datos tienen una distribución cercana a la normal.

Con relación al skewness, antes de empezar un test "t" es adecuado examinar el histograma para las diferencias en las variables.

Las variables con valores extremos o outliers deben ser cuidadosamente comprobados usando el procedimiento boxplots.

Hay procedimiento que se pueden usar para testar el supuesto de normalidad, especialmente en el procedimiento *Esploratory Data Analysis.*

Test pareados para muestras independientes

Testa la significancia de las diferencias entre las medias de dos muestras independientes; también nos provee de las estadísticas descriptivas para cada variable.

Por otro lado, nos proporciona un test de igualdad de varianzas y un coeficiente de intervalo al 95%, para la diferencia entre las medias de las variables.

Determinando los grupos

Generalmente, los grupos de dos muestras para un test "T" son fijos y las variables de grupos tienen un valor para cada grupo.

Pero hay ocasiones en las que se puede asignar estos valores para uno o dos grupos, sobre la base de una variable de escala.

Por ejemplo, consideremos las notas de los estudiantes en gramática y matemáticas.

Podríamos considerar la necesidad de realizar un test "T" sobre las calificaciones en gramática, dividiéndolos en dos grupos: los que tienen una calificación sobre el promedio en matemáticas y los que están por debajo

Con el procedimiento del test "T" para muestras independientes lo único que necesitamos es ese "punto de corte" que divide a los dos grupos que serán testados.

La bondad de este método es que el punto de corte puede ser modificado sin recurrir a crear otros grupos de variable.

Ejercicio 10.1
Un analista de una tienda de departamentos quiere evaluar los efectos de una carta de crédito provisional; para ello selecciona, aleatoriamente, a 500 clientes

La mitad de los clientes seleccionados reciben la carta de crédito con la reducción de la tasa de interés por los próximos 3 meses; la otra mitad recibe la tarjeta normal.

Los datos están en el archivo *creditpromo.sav*; usaremos el test "T" para dos muestras independientes y comparar los gastos de cada grupo

Menú → Analyze → Compare Means → Independent → spent during promotional period (como la variable de test) → *Type of mail insert received* (como la variable de grupo) → *Define groups* → escribimos 0 en la casilla del grupo 1 y escribimos 1 en la casilla de valor del grupo 2 → *Continue.* → *OK.*

Tabla de Resultados 10.4

Group Statistics

	Type of mail insert received	N	Mean	Std. Deviation	Std. Error Mean
$ spent during promotional period	Standard	250	1566,39	346,67	21,92
	New Promotion	250	1637,50	356,703	22,56

La Tabla de resultados anterior, consigna las medias de ambas muestras, como también sus desviaciones estándar y los errores estándar de las media.

Un primer vistazo a esa tabla, nos muestra que hay una diferencia de -71,11 entre la Media aritmética de las personas no tienen la promoción y las que sí tienen

Ahora nos corresponde establecer si esa diferencia entre las Medias es significativa. Para ello, analizamos la segunda tabla en la Pantalla de Resultados

La segunda tabla de la pantalla de resultados nos trae indicadores nuevos para nosotros; pero es muy extensa para trasladarla hasta aquí

Antes de comparar las medias de dos muestras, es necesario saber si las varianzas en la distribución de las medias son iguales o son diferentes

Para averiguarlo, vamos a recurrir al Test de Levene

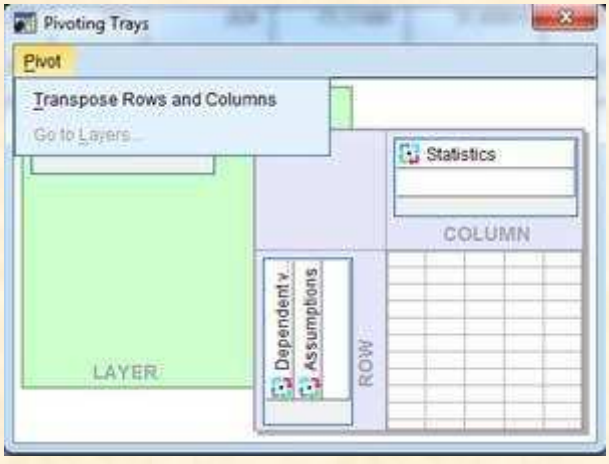
Dado que la pantalla es muy extensa, nos limitaremos a trasladar el segmento de la tabla grande que es de nuestro interés

Para ello, pivotearemos la tabla

Hacemos dos clicks seguidos en cualquier parte de la tabla
Aparecerá una nueva pantalla con un nuevo menú: *Pivote → pivote trays*

La pantalla de resultados nos da la siguiente bandeja

Bandeja 10.1



En la bandeja verde → click en Traspose Rows and Columns → cerrar
En la Pantalla de datos aparece una nueva tabla, esta vez resumida.

El Indicador F

En este caso, el Test F testa la hipótesis nula de que dos poblaciones con distribuciones normales tienen la misma varianza; luego veremos que tiene otros usos también.

Test de Levene

Nos ayuda a definir cuál de las dos hipótesis que vamos a plantear es valedera

Ho Las dos muestras tienen varianzas diferentes, una de la otra
H_1 Ambas muestras tienen la misma varianza

Si el nivel de Sig es mayor que el 5%, aceptamos la hipótesis alternativa
Esto es, aceptamos que ambas muestras tienen la misma varianza

Tabla de Resultados 10.5

			$ spent during promotional period	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		1,190	
	Sig.		,276	
t-test for Equality of Means	t		-2,260	-2,260
	df		498	497,595
	Sig. (2-tailed)		,024	,024
	Mean Difference		-71,11095	-71,11095
	Std. Error Difference		31,45914	31,45914
	95% Confidence Interval of the Difference	Lower	-132,91995	-132,92007
		Upper	-9,30196	-9,30183

El valor F es 1.19 y su nivel Sig = 0.276
Dado que el nivel Sig es mayor que el 5%, aceptamos la hipótesis alternativa

La hipótesis de la igualdad de varianzas entre ambas muestras se mantiene
Por lo que desechamos la información que nos proporciona la última columna

Luego de los estadísticos sobre la igualdad de varianzas, tenemos un test "t"

El Test "t" para la igualdad de las medias

Las hipótesis que deseamos testar en este caso es si las medias aritméticas de las dos muestras son iguales entre sí:

Ho = Las medias aritméticas de las muestras son iguales entre sí

H_1 = Las medias aritméticas de las muestras no son iguales entre sí

El estadístico "t" es: -2.26
El Sig de "t" es = 0.024 que es menor que el 5% que estamos usando

Este valor rechaza la hipótesis nula.
Esto es, las medias de ambas muestras no son iguales entre sí.

De todo esto, deducimos que la diferencia de -71,11$ es significativa entre el gasto de quienes tienen la carta de crédito promocional y los que no.

El SPSS obtiene la Diferencia de Medias substrayendo la media de la muestra del grupo 2 (el grupo promocional) de la media de la muestra del grupo 1 (no promocional)

Por eso es que el valor de -71,11 es negativo

La diferencia, en valore absolutos, de $71.11 dólares que incrementa el promedio del gasto de los beneficiarios con las tarjetas de crédito no se debe al azar.

Se debería considerar la posibilidad de extender la promoción a todos los clientes.

El Coeficiente de Intervalo de la diferencia entre ambas medias, al 95%, proporciona los límites inferior y superior entre los que se encuentra la verdadera media.

De todas las muestras de 500 clientes que pueden lograrse de la población, hay el 95% de posibilidades que la media verdadera se encuentre entre -132,91 y -9,3019.

Usando un punto de corte para definir las muestras

Hay una compañía que provee celulares y quiere estimar la propensión a abandonar los servicios de la firma, por parte de sus clientes.

Como referencia usa el punto de corte = 50 minutos
Los datos están el archivo *celular.save*

Usaremos el test "t" para muestras independientes y determinar si estos grupos tienen una diferencia significativa en el tiempo de uso de sus celulares, medido en minutos

> *Analyze > Compare Means > Independent → Avg monthly minutes como la variable a testar → Select Propensity to leave (como la variable de grupo) → Define Groups → Cut point → 50 en la casilla de cut point → Continue → OK.*

La tabla … nos muestra que el número de clientes por encima del umbral de quienes quieren abandonar el servicio de celulares de la compañía es 50.

El total de clientes cuyo promedio está por debajo del umbral para dejar el servicio es de 200 y el tiempo promedio de uso mensual de sus celulares es menor que 50 minutos

La Media del uso mensual de celulares de los que están por encima del umbral de 50 minutos, es 78 minutos mayor que la Media de quienes están por debajo del umbral

Tabla de Resultados 10.6

Group Statistics					
	Propensity to leave	N	Mean	Std. Deviation	Std. Error Mean
Avg monthly minutes	>50,00	50	224,79	37,73202	5,33611
	< 50,00	200	146,53	33,64859	2,37931

La segunda tabla nos mostrará si las diferencias en las Medias de ambas opciones son iguales entre sí o no.

Pero, ya lo sabemos, primero debemos recurrir al test de Levene para enterarnos si podemos asumir la igualdad de varianzas o no entre ambos conjuntos de datos

Sabemos que si no hay igualdad de varianzas, no será posible testar las diferencias
La tabla que necesitamos es muy extensa para traerla

Por lo que utilizaremos otra vez la función Pivote:

→doble ckick en cualquier lugar dentro de la tabla → Pivot en la nueva tabla
Aparece la bandeja; → pivot → trasspase rows anda coluns → cerrar

Tabla de Resultados 10.7

Independent Samples Test				
			Avg monthly minutes	
			Equal variances assumed	Equal variances not assumed
Levene's Test for Equality of Variances	F		,966	
	Sig.		,327	
t-test for Equality of Means	t		14,349	13,394
	df		248	69,742
	Sig. (2-tailed)		,000	,000
	Mean Difference		78,25625	78,25625
	Std. Error Difference		5,45394	5,84254
	95% Confidence Interval of the Difference	Lower	67,51430	66,60292
		Upper	88,99819	89,90958

El Test de Levene para contrastar la igualdad de varianzas tiene un valor F = 0.966
El nivel Sig es 0,327, por lo que aceptamos la igualdad de varianzas

El valor "t" para testar la igualdad de las Medias es 14,349 y su nivel Sig = 0.000
Por lo tanto, rechazamos la hipótesis de que las medias son iguales entre sí.

La diferencia en los tiempos del uso del celular es significativa

El "t" test proporciona una firme evidencia de que hay diferencia entre los minutos de uso de los celulares entre los que son más propensos a cambiar la firma que les provee el servicio y los que son menos propensos a cambiarla

Muestra que los que piensan abandonar el servicio tienen un promedio de uso mensual de 78 minutos más que el promedio de los que no piensan abandonarlo.

El Intervalo de Confianza tiene un valor mínimo de 67,51
Un valor máximo de 88.99

El intervalo de confianza sugiere que la diferencia real entre ambos grupos no será menor de 67 minutos, a pesar de que ese valor es superior al umbral de 50 minutos

La compañía debe tomar medidas para retener este grupo

Resumen

El test "t" es útil para comparar las medias de dos muestras independientes.
A veces es necesario establecer un punto de corte entre los dos grupos.

Como todos los tests "t" de muestras independientes, se asume que la media de cada muestra proviene de una población distribución sea cercana a la normal.

Esto se refiere, especialmente, al estadístico skewness

Por otra parte, las variables de valores extremos y outliers deben ser comprobadas; para ello contamos con el estadístico boxplot.

Hay otros procedimientos alternativos, bajo el supuesto de distribución normal en Exploratory Data Analysis.

Si se tiene que testar más de dos grupos de variables, acudimos al procedimiento One-Way ANOVA que es el tema del próximo capítulo.

11
ONE WAY ANOVA

ANOVA deriva de las iniciales en inglés "Analysis of Variance"

Conceptos

El Test "t" es muy útil para muestras pequeñas y para establecer la diferencia entre las Medias Aritméticas, pero solamente se refiere a la comparación de dos variables

ANOVA amplía el rango del "t" test para determinar si las medias, no sólo de dos grupos, sino de varios son o no iguales.

También ofrece niveles de grupo para la variable dependiente, un test sobre igualdad de varianzas, una gráfica de medias de grupos.

Por otra parte, tests de rangos, comparaciones múltiples de grupos pareados y contrastes para describir la naturaleza de las diferencias de grupo.

Test de Varianzas

Un importante paso en el análisis de varianza es el test de la validez de los supuestos; por ejemplo, ANOVA supone que la varianza de grupos es equivalente.

Si deseamos analizar la variabilidad de una variable numérica debido al comportamiento de otra, pero de tipo categórica, ésta última recibe el nombre *factor.*

En este caso, estaremos hablando del análisis de la varianza para un solo factor, aunque este factor, a pesar de ser solo uno, puede tener varios niveles.

Un ejemplo sería el análisis de los promedios del ingreso de dos grupos independientes de funcionarios, de acuerdo con el grado de educación que tienen.

La variable dependiente, cuantitativa, será el Ingreso.
El grado de educación será el *factor.*

Primaria, secundaria, universitario y post grado serán niveles del factor *educación*

ANOVA compara las medias de ambos grupos para establecer si hay o no diferencia y, en su caso, el grado de diferencia.

Estos niveles pueden ser fijos, tales como los que acabamos de describir.
Aleatorios; v.g, si se escoge al azar las empresas que participarán en el estudio.

Si los resultados proporcionados por las muestras no tienen errores sistemáticos, sus medias no serán significativamente diferentes entre sí.

La dispersión que pudiera notarse se deberá a errores aleatorios que representarán la dispersión de los elementos de una muestra.

Fuentes de variación
Habrá dos fuentes de variación cuando comparamos dos muestras:

La que se debe a la diferencia de los estadísticos que separan ambas medias.
La que se debe a los errores dentro de cada muestra.

La Suma Total de Cuadrados
SS_T: Suma total de estas variaciones

Se denomina *Suma Total de Cuadrados*.
Esta Suma Total se descompone en dos sumas de cuadrados.

ST_T: Suma de las diferencias al cuadrado de cada resultado individual con relación a la Media de todos los resultados; representa la variación total de los datos.

SS_I; Mide las desviaciones entre los resultados individuales de cada muestra con relación a su propia Media; es una medida de dispersión dentro de cada muestra.

SS_E; Mide las diferencias entre los resultados medios de cada muestra y el resultado medio global de todas las muestras; de este modo se tiene:

$$ST_T = SS_I + SS_E \qquad (11.1)$$

Si se divide SS_I por los correspondientes grados de libertad se obtiene la Media Cuadrática (Mean Square) dentro de cada muestra (MS_I)

De esta manera, $ST_T = SS_I + SS_E$

¿Por qué "Diferencia de cuadrados"?
Tomemos la siguiente serie de números: 1, 2, 3, 4, 5
La Media aritmética de esa serie será: (1 + 2 + 3 + 4 + 5)/5 = 3

Ahora bien, tomemos la suma de cada elemento individual con relación a la media:
(1 – 3), (2 – 3), (3 – 3), (4 – 3), (5 – 3) y las sumemos, -2 + (-1) + 0 + 1 + 2 = 0

El resultado es cero, debido a que los valores negativos de las diferencias se contrapesan con los resultados positivos, tal como lo vimos en capítulo anterior.

Para evitar que eso suceda, se suma los cuadrados de las diferencias, puesto que los cuadrados de números negativos o positivos siempre serán positivos.

De esta manera tendremos:

$$(1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2 = 1 + 1 + 0 + 1 + 4 = 7$$

Éste resultado es la *Suma de Cuadrados*, que se usa para medir las diferencias.
Ahora nos enfocaremos en el modelo ANOVA de una sola vía (One way ANOVA)

One Way of Variance (ANOVA)

One Way ANOVA nos permite testar la hipótesis de que las varianzas de dos o más grupos son o no son significativamente diferentes unas de las otras.

Ejemplo

Un gerente de ventas quiere determinar el número óptimo de días de entrenamiento para algunos empleados nuevos.

Hasta el momento tiene los datos del desempeño de tres grupos:
Empleados con uno, dos y tres días de entrenamiento, respectivamente.

Para empezar el análisis de esos datos traemos el archivo *salesperformance.sav*
Antes de correr el ANOVA es necesario graficar las medias y los errores típicos.

Menú principal → Gráphics → Chart Builder → galería →
En la parte inferior aparecen varias opciones de gráficas

Con el mouse apuntamos a cada una hasta que identifiquemos el ícono Simple Error Bar

Con el mouse, arrastramos el ícono a la pantalla blanca
Con el mouse también arrastramos *Score on training exam* en el eje de las "Y"

Con el botón derecho click *Sales training group* y elegimos *Nomina*l
Con el mouse arrastramos *Sales training group* al eje de las X's

Con el mouse, arrastrar el ícono *Simple Error Bar en la pantalla blanca.*
Luego arrastramos hasta el eje de las Y la variable *Score on training exam.*

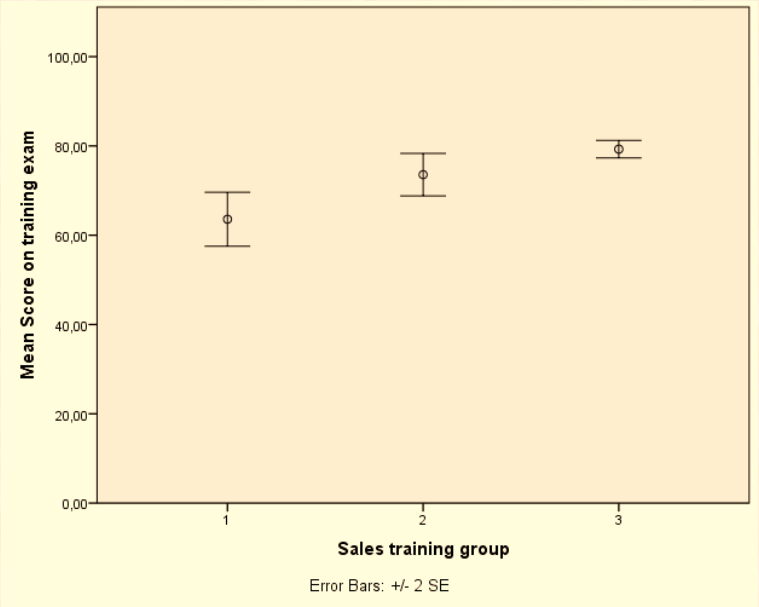
Con el botón derecho → *Sales training groups → Nominal*
Arrastrar la variable Sales training groups al eje de las X → Click *element properties*

Barra de Error → Standard Error → Apply → OK en el chart builder

La gráfica 11.1 nos muestra que el desempeño de los funcionarios aumenta a medida que se incrementa el número de días de entrenamiento.

Pero, al mismo tiempo, la variación en el desempeño se reduce, es decir, al parecer las varianzas en cada grupo no son iguales

Diagrama 11.1



Sabemos que ANOVA asume igualdad de varianzas en todos los grupos, pero ese supuesto puede que no se cumpla para este caso.

Es necesario testar el supuesto de igualdad de varianzas

Analyse → Compare Means → One-Way ANOVA
Score on training como la variable dependiente y *Sales training* (variable factor)
→ Options → Click Homogeneity of variance test → Continue → OK

Tabla de Resultados 11.1

Descriptives

Score on training exam

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	20	63,57	13,51	3,0206	57,2576	69,9020	32,68	86,66
2	20	73,56	10,61	2,3722	68,6025	78,5328	47,56	89,65
3	20	79,27	4,46	,9855	77,2165	81,3420	71,77	89,69
Total	60	72,14	12,00	1,5496	69,0415	75,2430	32,68	89,69

Las desviaciones estándar y los errores estándar decrecen a medida que los cursos tienen una duración mayor en días

Esto nos confirma que a medida que el tiempo de entrenamiento aumenta, la variación en el desempeño de los participantes disminuye.

La segunda se refiere a la homogeneidad de la varianza en los exámenes después de los días de entrenamiento en cada caso

La prueba de Levene tiene un valor de 4,6370 y un nivel Sig 0.014 que es menor que el 5% de error que acordado, por lo tanto, rechazamos la hipótesis de varianzas iguales.

Tabla de Resultados 11.2

Test of Homogeneity of Variances			
Score on training exam			
Levene Statistic	df1	df2	Sig.
4,6370	2	57	,014

Heterocedasticidad

Nos indica que las diferentes muestras a ser comparadas tienen varianzas diferentes.

Algunos autores estiman que ése es un gran impedimento para comparar medias y desviaciones estándar.

Implicaciones de la no igualdad de varianzas

N. George Mankiw, en su libro "*A Quick Refresher Course in Macroeconomics, Journal of Economics Literature noviembre 1990",* afirma al respecto:

"La heterocedasticidad jamás ha sido una razón para desechar un modelo que de otra forma sería adecuado".

Por su parte, John Fox en su obra "Aplied Regression Análysis", pg. 306 afirma:

"… *una varianza de error desigual vale la pena corregirlo sólo cuando el problema resulta severo."*

Supongamos que se nos presentaran muestras con diferentes varianzas.

En ese caso, ceptaríamos las dos opiniones citadas y decidiríamos que la diferencia de las varianzas no es algo que nos impida seguir analizando los datos respectivos.

Otra aplicación de One Way ANOVA

Una firma electrónica desarrolla un nuevo reproductor de DVD; se lanzó un prototipo y se recolectaron los datos de dos grupos divididos de acuerdo con la edad.

ANOVA puede estableces si los clientes de grupos de edades distintas reaccionaron de maneras diferentes ante el nuevo diseño; traemos el archivo *dvdplayer.sav*

> *Comparar Medias → One Way ANOVA → Total DVD Assessment, (dependiente) → Age group (Factor) → Opciones → Prueba de homogeneidad de las muestras → Gráfico de las Medias → Continuar → OK.*

La primera tabla es el test de Levene sobre homogeneidad

Tabla de Resultados 11.3

Test of Homogeneity of Variances			
Total DVD assessment			
Levene Statistic	df1	df2	Sig.
1,292	5	62	,279

El nivel Sig del test Levene es 0.279 que es mayor que 0.05; por lo tanto, aceptamos la hipótesis nula: en este caso las varianzas son iguales. Vamos a la tabla 11.5.

La segunda tabla nos trae los indicadores de ANOVA

Tabla de Resultados 11.4

ANOVA					
Total DVD assessment					
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1294,481	5	258,896	6,993	,000
Within Groups	2295,532	62	37,025		
Total	3590,013	67			

El Estadístico F

Ahora concretamos el análisis que hicimos anteriormente sobre el estadístico F.

La Prueba F contrasta *la Hipótesis nula* de que las distribuciones normales de las poblaciones tienen la misma varianza.

El nivel de Sig del test F es 0,000, por lo que rechazamos la hipótesis de que las medias entre los grupos por edades son iguales.

La suma Total (ST_T en nuestra simbología) suma las diferencias al cuadrado de cada resultado individual con relación a la Media de todos los resultados.

Representa la variación total de los datos.

La Suma de Cuadrados *Intra grupos* mide las diferencias entre los valores de cada muestra con relación a su propia Media.

Es una medida de dispersión dentro de cada muestra
Es el error en ANOVA.

La Suma de Cuadrados *Inter grupos* es la Media de las diferencias entre los elementos de las muestras con relación a la Media general.

En realidad la *Suma de Cuadrados Inter grupos* es la que importa para los análisis.
En el cuadro ANOVA vemos que la *Suma de Cuadrados Inter Grupo* es 1294,48.

Si se divide la Suma de cuadrados Inter-grupos entre los correspondientes grados de libertad, se obtiene la Media Cuadrática en cada caso (Mean Square)

En este caso, la Media Cuadrática es 258,896.

La Media Cuadrática para la suma de cuadrados Intra-grupos también resulta de la división de la Suma de Cuadrados Intra-grupos entre sus grados de libertad.

La Media Cuadrática correspondiente es 37.025.
La Suma Total de Cuadrados es 1294,481 + 2295,532 = 3590.013

Los grados de libertad para el total es igual al número de todos los elementos que fueron tomados en cuenta en las muestras menos 1; en nuestro caso es 67.

Los grados de libertad para la Suma de cuadrados inter-grupos es igual a K - 1 donde K es el número de muestras; en nuestro caso, 5.

Los grados de libertad para la Suma de cuadrados Intra-grupos es igual a N – K = 62.
Ahora nos toca interpretar la Prueba F.

El valor de F se calcula dividiendo la Media cuadrática Inter-grupos sentre la Media cuadrática intra grupos

Esto es, la variabilidad explicada sobre la variabilidad no explicada

258,8960/37,0250 = 6,9930

El resultado es el mismo que tiene el valor del estadístico F en la tabla anterior….

Sobre este resultado es posible afirmar que los clientes de grupos de edades distintas reaccionaron de maneras diferentes ante el nuevo diseño.

Ya sabemos que las actitudes de las personas en los diferentes grupos, no son iguales Ahora aprenderemos acerca de la estructura de las diferencias

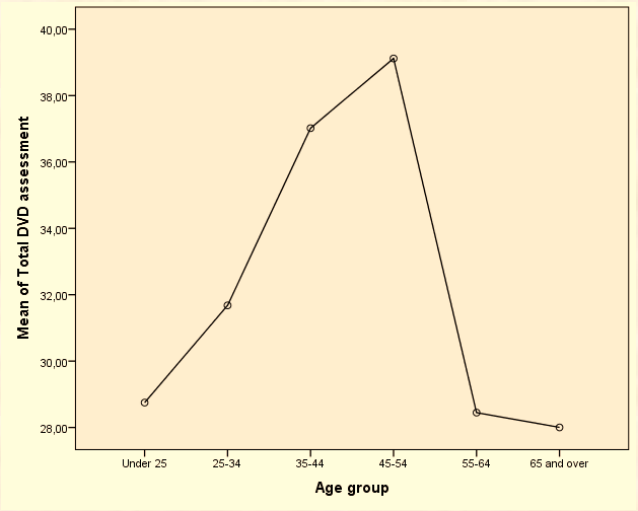
Contrastes entre las medias de cada grupo

Una vez que sabemos que hay diferencia entre los grupos de entrenamiento, deseamos conocer la estructura de las diferencias.

Para ello recurrimos a la gráfica de las Medias entre los grupos.

En la gráfic… vemos que los participantes entre 35 y 54 años de edad son los que aceptaron con mayor receptividad el nuevo reproductor de DVD

Gráfica 11.1



Comparaciones de pares múltiples

El método que vamos a utilizar es conocido como comparaciones de pares de medias aritméticas cuando los grupos son múltiples.

De este modo, ANOVA nos ayuda a comparar las medias de un grupo con la media de cualquier otro grupo, ciuando los grpos son más de dos.

Ejemplo

Un gerente de ventas ha analizado los datos referidos a un programa de actualización usando el procedimiento One-Way ANOV.

Aunque se encontraron diferencias significativas de grupo, no se cuenta con una hipó-tesis previa acerca de la manera cómo difieren los grupos.

Después de terminadas las clases de actualización todos los participantes dieron un examen para establecer el grado de aprovechamiento.

De este modo, decide comparar cada grupo con otro, individualmente, sobre la base de los datos que se archivaron en *salesperformance.sav.*

Empezamos el análisis pertinente.

Analyze > *Compare Means* > *One-Way ANOVA* →*Score on training exam* (como variable dependiente) → *Sales training group* (Como variable factor) → *Post Hoc.*

La Caja de Post Hoc se divide en dos partes; la primera asume grupos con varianzas iguales; la segunda no asume igualdad de varianzas

Sesupone que el Test Leven ha establecido que las varianzas de los diferentes grupos son dignificativamente diferentes, vamos a escoger esa opción que está en la parte inferior

Hacemos click en el botón Tamhane's test → Continue → OK

Tabla de Resultados 11.5

Multiple Comparisons

Dependent Variable: Score on training exam

Tamhane

(I) Sales training group (II)	(J) Sales training group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-9,98789*	3,84	,040	-19,60	-,3705
	3	-15,69947*	3,18	,000	-23,82	-7,5198
2	1	9,98789*	3,84	,040	,37	19,6053
	3	-5,71158	2,57	,102	-12,28	,8539
3	1	15,69947*	3,18	,000	7,52	23,8792
	2	5,71158	2,57	,102	-,85	12,2771

*. The mean difference is significant at the 0.05 level.

En la tabla anterior la información sobre los resultados de los exámenes se ha dividido en grupos, de acuerdo al número de días que asistieron a las clases de actualización.

La letra (J) designa a los otros grupos con los cuales el grupo (I) se compara

En el lado izquierdo, el grupo que tuvo un solo día de actualización se compara con los que tuvieron 2 y 3 días respectivamente

La primera fila de la segunda columna muestra las diferencia en las medias entre el primer grupo y los otros dos (-9,98789 y -15,69947, respectivamente)

Así, el promedio de las diferencias de las medias en los exámenes que obtuvo el grupo de quienes tuvieron sólo un día de clases con el grupo que tuvo 2 días es -9.98789.

La diferencia en el promedio de exámenes entre los participantes del mismo grupo que tuvo un solo día de clases de actualización con el que tuvo 3 días, es de -15,6997

Dado que las diferencias en ambos casos es negativa, deducimos que el desempeño de los que asistieron un día a las clases fue menor que el de los otros dos grupos.

El número 2, en azul, representa al grupo que asistió 2 días a las clases; la diferencia de las calificaciones del examen con relación al grupo 1 se repite, por supuesto.

Pero esta vez con signo positivo, dado que el rendimiento de los participantes del grupo 2 fue mayor.

La diferencia de las calificaciones del grupo 2 con las del grupo que asistió 3 días a las clases es negativa, -5,71158

La comparación del grupo 3 tiene el mismo procedimiento.
El nivel Sig para las diferencias de las Medias entre el grupo 1 y el 2 es Sig =0.040

En razón de que el nivel Sig es menor que el 5% deducimos que hay una diferencia estadísticamente significativa entre los que asistieron un día y los que fueron dos días

El nivel Sig de contraste entre el grupo que asistió dos días y el de los que asistieron tres días al programa de actualización es 10.2

Este valor Sig es mayor que el 5%

Por lo que deducimos que no hay una diferencia estadísticamente significativa entre los grupos que asistieron 2 días, por una parte, y el grupo que tuvo 3 días de actualización.

Resumen

Por lo analizado hasta ahora, deducimos que con el procedimiento One-Way ANOVA podemos realizar varias operaciones importantes:

Validar o no el supuesto de la igualdad de varianzas en varios grupos.
Obtener todos los resultados que ANOVA pone a disposición

Visualizar los promedios de cada grupo.
Realizar contrastes personalizados para hipótesis específicas.

Comparar cada media con todas y cada una de las demás, ya sea asumiendo igualdad de varianzas o no.

Otros Procedimientos similares

Vimos que One-Way ANOVA es usado para testar hipótesis sobre la igualdad de las medias de varios grupos, no sólo de dos.

También se puede usar al procedimiento Means para obtener one-way análisis de varianza con un test de linearialidad

Si es necesario usar factores múltiples, debe recurrirse al procedimiento GLM Univariate para obtener two-way ANOVAde la covarianza y más.

Eso es precisamente lo que haremos en el próximo capítulo.

12
EL PROCEDIMIENTO GLM

El procedimiento GLM Univariado nos permite ver la relación de una variable dependiente numérica con otras variables categóricas y predictores numéricos.

Se basa en al procedimiento del Modelo General Lineal en el que los factores y covariables tienen alguna relación lineal con la variable dependiente.

Factores

Los factores del modelo son variables categóricas o de escala; cada nivel de un factor dado puede tener un efecto lineal en el valor de la variable dependiente.

La Variable dependiente es cuantitativa.
Los factores fijos son las variables cuyos valores de interés se presentan en los datos.

Los factores aleatorios *(Random-effect factors*) son variables cuyos valores pueden ser considerados como una muestra aleatoria de una población grande de valores.

Por ejemplo, una tienda puede estar interesada en determinar cuáles factores, de la variedad existente afectan las compras.

Puede ser que le interese determinar si el Género y los Estilos de compra mensual que realizan los clientes en una cadena de tiendas, influyen sobre las compras

En este caso, se tomaría como factores el género y el estilo de compras.

Estas últimas podrían clasificarse como compras dos veces por semana, una vez por semana y otros similares.

Por otro lado, utilizar más de un factor permite identificar la influencia, no sólo de cada factor sobre la variable dependiente, sino la interacción de los factores.

De este modo, en un análisis de dos factores, v.g, hay tres efectos que nos interesan: la de cada factor (dos efectos en este caso) y el efecto de la interacción.

El fin principal es apreciar la importancia de las variaciones debidas a una diferencia real entre dos muestras con relación a la diferencia que surge por simple azar.

Por eso es que los teóricos de la Estadística han se han encontrado con la necesidad de descomponer la suma total de cuadrados en dos vertientes.

Ya lo vimos en el capítulo anterior, al iniciar el ANOVA: la vertiente que viene de la variación real entre las muestras y la vertiente que proviene de azar.

Si la Suma de cuadrados debida a la verdadera diferencia entre las muestras supera a la que viene del azar, ese cociente será mayor que 1.

Entonces podremos afirmar que hay una diferencia real entre las muestras.

GLM Univariate para realizar un Two-Factor Analysis of Variance

Una tienda de abarrotes está interesada en los efectos de cinco diferentes tipos de cupones sobre los gastos de los clientes; archivo *grocery_1month.sav*

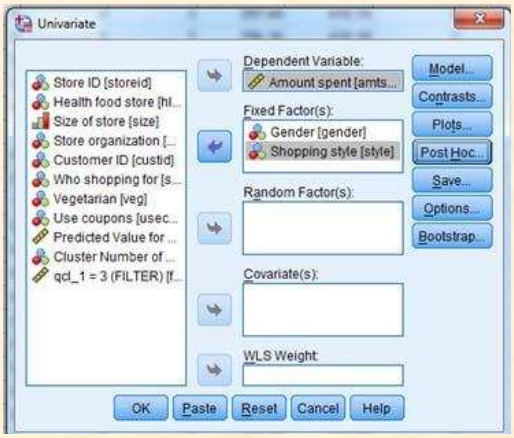
Usaremos el procedimiento GLM Univariate para realizar un ANOVA de dos factores, o lo que ese lo mismo, un procedimiento Two-Way ANOVA

Menú → Analyze → General Model → Univariate

→ Amount spent (Como la variable dependiente)
→Gender and Shopping style (Factores fijos)

Hasta el momento, lo realizado debe lucir tal como se muestra en la Caja 12.1

Caja 12.1



→ *Plots* → *style* en el eje de las X → *Gender* en la casilla "separate lines variable" → *Add* → *Continue* → *Post Hoc* en la caja de diálogo GLM Univariate

→ *style* (como la variable para la que se realizará el test post hoc) → *Tukey (*en la casilla "Equal Variances Assumed group → *Continue* → *Options*

→ *gender*style* como el elemento para el que se desplegará las medias

→*Descriptive statistics* → *Homogeneity tests* → *Estimates of effect size* → *Spread vs. level plot* en el despliegue del grupo → *Continue* → *OK*

Tabla de Resultados 12.1

Descriptive Statistics				
Dependent Variable: Amount spent				
Gender	Shopping style	Mean	Std. Deviation	N
Male	Biweekly; in bulk	413,06	90,86574	35
	Weekly; similar items	440,96	98,23860	120
	Often; what's on sale	407,77	69,33334	30
	Total	430,30	93,47877	185
Female	Biweekly; in bulk	343,98	100,47207	35
	Weekly; similar items	361,72	90,46076	102
	Often; what's on sale	405,72	80,57058	29
	Total	365,67	92,64058	166
Total	Biweekly; in bulk	378,52	101,25839	70
	Weekly; similar items	404,55	102,48440	222
	Often; what's on sale	406,77	74,42114	59
	Total	399,73	98,40821	351

En la pantalla de datos se despliegan tres tablas, de las que traemos primero la que etiquetamos como tabla

La tabla despliega estadísticas descriptivas para cada una de las combinaciones de factores en el modelo; podría detectarse un efecto debido a shopping style.

En promedio, las compras dos veces por semana (biweekly) los clientes gastan $378.52, mientras que los clientes que compran una vez por semana, gastan $404.55

Los clientes que compran a menudo (often) gastan $406.76 en cada compra.

El efecto Gender; en el promedio los hombres gastan $430.30 en comparación con las compras que realizan las mujeres, que es de $365.66 en promedio.

Hay un efecto de interacción entre "Gender" y "Shopping Style"
Pues las diferencias de sus medias en el total del gasto clasificado por shopping style varían entre los géneros (Genders)

Los clientes varones que compran dos veces por semana gastan más en cada compra (413.0657) que los clientes clasificados en "often = liquidación" (407.77)

Sin embargo, la tendencia se revierte para las compras realizadas por mujeres en las categorías dos veces por semana ($ 343.98) y often = liquidación ($405.73)

La columna N de la tabla muestra que las celdas tienen diferentes tamaños.
La mayor parte de los clientes prefiere realizar sus compras semanalmente.

Las desviaciones standard parecen relativamente homogéneas, para asegurarnos traemos la tabla correspondiente al test Levene

Test de Levene

La tabla… testa la hipótesis nula de que la varianza del término de error es constante en todas las celdas, definidas por la combinación de los niveles de factor

Dado que el valor Sig = 0.330, es mayor que 0.10, por lo que aceptamos la hipótesis nula de que no hay diferencias entre las varianzas del total de las celdas

Las pequeñas diferencias en las desviaciones estándar de los grupos que observamos en la tabla de estadísticas descriptivas, son variaciones debidas al azar.

La gráfica de puntos spread vs. level registra las medias aritméticas de cada celda y las respectivas desviaciones estándar de la tabla 12.2

La gráfica 12.1 nos muestra un cuadro visual del supuesto de igualdad de varianzas

Tabla de Resultados 12.2

Levene's Test of Equality of Error Variances[a]			
Dependent Variable: Amount spent			
F	df1	df2	Sig.
1,157	5	345	,330

Incluye información para averiguar si las violaciones al supuesto se deben a la relación entre las medias de las celdas y sus desviaciones estándar.

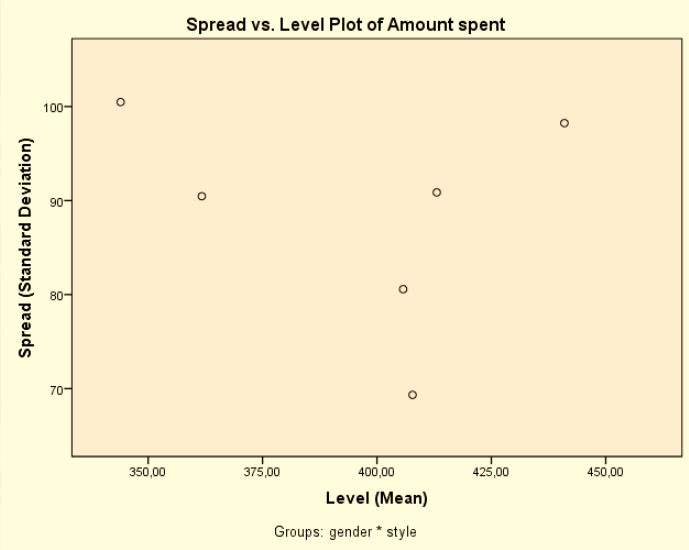
La dispersión de los puntos de la gráfica 12.1 muestra que no hay una relación estadisticamente significativa entre las medias y las desviaciones estándar.

La Tabla 12.3 testa la importancia de un factor, aunque no nos indica la manera en que cada factor varía

La gráfica 12 muestra la relación entre la Media y la Desviación Estándar.

Incluye información para averiguar si las violaciones al supuesto se deben a la relación entre las medias de las celdas y sus desviaciones estándar.

Gráfica 12.1



La dispersión de los puntos de la gráfica 12.1 muestra que no hay una relación estadisticamente significativa entre las medias y las desviaciones estándar.

El test post hoc muestra las diferencias en el modelo de predicción de las medias para cada par de niveles de factor; la Tabla de Resultados 12.3 muestra las diferencias de estilos de compra entre hombres y mujeres

Tabla de Resultados 12.3

Gender * Shopping style					
Dependent Variable: Amount spent					
Gender	Shopping style	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Male	Biweekly; in bulk	413,07	15,55	382,48	443,65
	Weekly; similar items	440,96	8,39	424,45	457,48
	Often; what's on sale	407,77	16,79	374,74	440,81
Female	Biweekly; in bulk	343,98	15,55	313,39	374,56
	Weekly; similar items	361,72	9,11	343,80	379,64
	Often; what's on sale	405,73	17,08	372,12	439,33

Por lo general, los hombres tienen un gasto mayor en las compras. La Tabla 12.3 testa la importancia de un factor, aunque no nos indica la manera en que cada factor varía.

El test post hoc muestra las diferencias en el modelo de predicción de las medias para cada par de niveles de factor.

Tabla de Resultados 12.4

Multiple Comparisons						
Dependent Variable: Amount spent Tukey HSD						
(I) Shopping style	(J) Shopping style	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Biweekly; in bulk	Weekly; similar items	-26,03	12,61	,099	-55,72	3,65
	Often; what's on sale	-28,23	16,26	,193	-66,52	10,02
Weekly; similar items	Biweekly; in bulk	26,03	12,61	,099	-3,65	55,72
	Often; what's on sale	-2,21	13,47	,985	-33,93	29,51
Often; what's on sale	Biweekly; in bulk	28,25	16,26	,193	-10,02	66,52
	Weekly; similar items	2,21	13,47	,985	-29,51	33,93
Based on observed means. The error term is Mean Square (Error) = 8463,939.						

En la tabla 12.4 vemos la información que necesitamos

Al igual que en el anterior capítulo, la variable que se desea comparar con las otras es la que se etiqueta con (I) las variable con las que compara son las (J) en cada caso

En la primera fila de la primera columna se compara Shopping Style = (I) con las otras dos modalidades de compra: *Weekly similar ítems* y *Often, wht's on sale*.

La diferencia entre las media de quienes compran bisemanalmente (Biweekly) y la media de quienes compran una vez por semana (Weekly) es -26.03

En la tabla 12.4 todos los valores Sig son superiores a 0.05.

De allí deducimos que no habría diferencias significativas entre los hábitos de compra establecidos: "biweekly", "weekly", or "often" por parte de los clientes.

Las medias aritméticas de los grupos se presentan en sub conjuntos homogéneos y fueron deducidas las medias observadas.

Los sub conjuntos homogéneos de la tabla traen los resultados de los test post hoc.

En la columna de sub conjuntos (subtes) los niveles de factor que no tienen diferentes efectos significativos.

En la tabla 12.5, el primer sub conjunto contiene las modalidades de compra que tienen los clientes: "biweekly", "weekly", and "often".

Es el único subconjunto, puesto que allí están representados todos los clientes.

Tabla de Resultados 12.5

Amount spent		
Tukey HSD		
Shopping style	N	Subset
		1
Biweekly; in bulk	70	378,52
Weekly; similar items	222	404,55
Often; what's on sale	59	406,77
Sig.		,12

Las medias aritméticas de los grupos se presentan en sub conjuntos homogéneos y fueron deducidas las medias observadas.

Los sub conjuntos homogéneos de la tabla traen los resultados de los test post hoc.

En la columna de sub conjuntos (subtes) los niveles de factor que no tienen diferentes efectos significativos.

En la tabla 12.4, el primer sub conjunto contiene las modalidades de compra que tienen los clientes: "biweekly", "weekly", and "often".

Es el único subconjunto, puesto que allí están representados todos los clientes.

El test post hoc sugiere que no es necesario incitar a los clientes para que compren más a menudo de lo usual porque el gasto no aumentará significativamente.

Los resultados del test post hoc no toman en cuenta los niveles de otros factores.

De este modo, ignoran la posibilidad de un efecto interactivo de Gender y la tabla de estadísticos descriptivos

Para tener una mejor información veremos los estimados de las medias marginales y comprobar cómo pueden cambiar nuestras conclusiones

Para ello trasladamos la tabla 12.6 de la Pantalla de Resultados.

Ubicamos las medias marginales del modelo y sus desviaciones estándar de los gastos con relación a las combinaciones de los factores Gender y Shopping style.

La información que nos proporciona la tabla 12.6 es muy importante para explorar los posibles efectos de interacción entre los dos factores.

Tabla 12.6
Gender * Shopping style

Dependent Variable: Amount spent					
Gender	Shopping style	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Male	Biweekly; in bulk	413,07	15,55	382,47	443,65
	Weekly; similar items	440,96	8,39	424,44	457,48
	Often; what's on sale	407,77	16,79	374,73	440,81
Female	Biweekly; in bulk	343,97	15,55	313,39	374,56
	Weekly; similar items	361,720	9,11	343,80	379,63
	Often; what's on sale	405,72	17,08	372,12	439,32

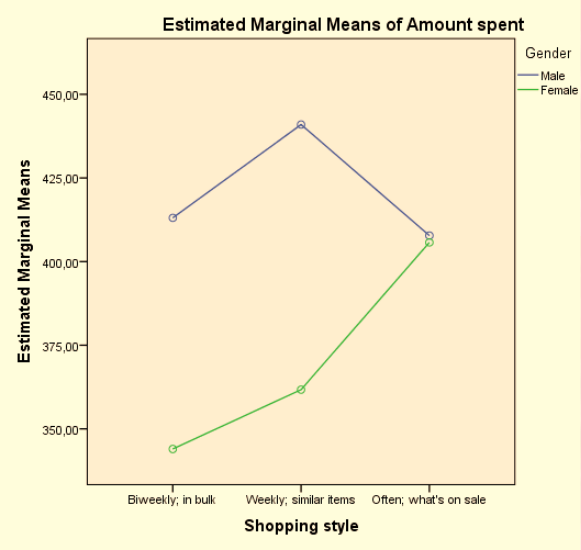
Se espera que un cliente masculino que realiza sus compras weekly, gaste $440.96, mientras que esperaríamos que otro cliente que compra "Often" gaste $407.77

También se espera que una mujer que compra "weekly" gaste $ 361.72 mientras que otra, de la modalidad "Often", gaste $405.72

De este modo concluimos que hay una diferencia significativa entre las modalidades de compra "weekly" y "often", la que depende del género del cliente.

Esto nos sugiere que existe un efecto de interacción entre Gender y Shopping style.

Gráfica 12.2



Si no hubiera interacción esperaríamos que la diferencia entre shopping styles permaneciera constante entre clientes masculinos y femeninos.

La interacción puede ser detectada fácilmente en los gráficos de perfil, tabla 12.2
En el eje horizontal se registra los niveles del factor Shoping style

El gráfico diseña líneas separadas para cada nivel de Gender.
Si no hay efectos de interacción, las líneas en el gráfico serían paralelas

En cambio, la diferencia de los gastos realizados entre los clientes que compran bajo las modalidades de "weekly" y "often" es mayor para la clientela femenina.

Lo deducimos porque las líneas para la clientela femenina tienen pendiente positiva (hacia arriba) y las de la clientela masculina, pendiente negativa (hacia abajo)

Hay un efecto de interacción fuerte que no es probable que se deba al azar.

Pero, para asegurarse, se puede comprobar el grado de significación con los tests de los efectos "between-subjects"

Para este propósito, copiamos de la Pantalla de Resultados la tabla que llamaremos Tabla 12.7 y que es una tabla de ANOVA

Tabla 12.7: Dependent Variable: Amount Spent

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	469403[a]	5	93880,59	11,09	,000
Intercept	39359636,38	1	39359636,38	4650,27	,000
gender	158037,44	1	158037,44	18,67	,000
style	33506,21	2	16753,10	1,98	,140
gender * style	69858,32	2	34929,16	4,12	,017
Error	2920058,82	345	8463,94		
Total	59475118,44	351			
Corrected Total	3389461,820	350			

Cada término en el modelo, más el modelo como un todo es testado acerca de su capacidad de tomar en cuenta la variación en la variable dependiente

Las etiquetas de las variables no se registran en la tabla

El valor Sig para cada término, excepto para Style, es menos que 0.05; así concluimos en que cada término, excepto Style es estadísticamente significativo

No tomamos en cuenta la última columna de la tabla que figura en la Pantalla de resultados, puesto que aún no hemos conceptualizado el estadístico Eta y sus variantes

Resumen

En este ejemplo vimos que los tests post hoc no revelan diferencias significativas entre los clientes que compran "weekly" y los que compran en la modalidad "Often"

Sin embargo, los estimados de las medias marginales y los gráficos de perfil revelaron una interacción entre los dos factores.

Lo que sugirió que la clientela masculina que compra una vez a la semana (weekly) es más rentable que los que compran en la modalidad Often.

Esa tendencia se revierte para la clientela femenina; la significancia del efecto interacción fue confirmada por los resultados de la tabla ANOVA.

El uso del GLM Univariate para realizar un Analysis of Covariance.

Se desea testar la bondad de un programa de trabajo para ayudar a la gente en sus esfuerzos de buscar trabajo, para lo cual tenemos una variable de control

Es el salario que cada participante tenía antes de ingresar al programa.

Del total de los participantes, algunos fueron elegidos al azar para ingresar al programa, mientras que otros quedaron excluidos, compararemos los dos grupos.

Antes de correr el programa definiremos un nuevo término

La Covariada

Es una variable secundaria que afecta la relación entre una variable dependiente y otras variables independientes

Los datos están en el archivo *workprog.sav*

Usaremos el procedimiento GLM Univariado para realizar una análisis de covarianza (ANCOVA) sobre los ingresos del programa.

Un supuesto extra de ANCOVA es que no hay una interacción significativa entre las covariadas del factor, así iniciamos el modelo con un término de interacción

Analyze → *General Linear Model* → *Univariate* → *Reset* → *Income after the program* (como la variable dependiente) → *Program status* (Como el factor fijo) → *Income before the program* (como la covariada) → *Model*

Elegimos *Custom* (como el tipo de modelo)

En la lista de factores y covariadas, a la izquierda hacemos click, manchando las dos variables: prog and incbef

Elegimos Build Term(s) drop-down list buscamos Main effects y con la fleche introducimos las dos variables a la pantalla de la derecha.

Otra vez volvemos a la lista de variable para manchar con click a las dos variables.
Esta vez vamos a Build Term(s) drop-down list pero buscamos Interaction

Click en la flecha y en la apantalla de la derecha se registra incbef*prog
Continue → Options in the GLM Univariate dialog box.

De la parte inferior (Diplaye) elegimos Estimates of effect size → Continue → OK
En la Tabla 12.8 se registran los estadísticos que hemos convocado.

Los significados de las abreviaciones son los siguientes:

prog = el programa de ayuda
incbef = salario antes de ingresar al programa
prog*incbef = la interacción de las dos variables

Tabla 12.9

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12295,033[a]	3	4098,344	429,755	,000
Intercept	131,271	1	131,271	13,765	,000
prog	106,795	1	106,795	11,199	,001
incbef	7152,586	1	7152,586	750,025	,000
prog * incbef	4,292	1	4,292	,450	,502
Error	9498,318	996	9,536		
Total	297121,000	1000			
Corrected Total	21793,351	999			

Se ha cortado la columna sobre Partial Eta Squared, estadístico que veremos en otro capítulo, con mayor detalle

La llamada al pie del cuadro se refiere al Coeficiente de correlación, cuyo significado será muy importante en el segundo tomo de esta obra: Econometría Aplicada.

Lo primero que vemos es el valor Sig = 0.502 de la interacción prog*incbef es mayor que 0.05, lo que nos indica que la interacción no es importante.

Volvamos al cuadro de diálogo GLM Univariate → Model → Full factorial → Continue → Options in the GLM Univariate dialog box.

Tabla 12.9 Descriptive Statistics			
Dependent Variable: Income after the program			
Program status	Mean	Std. Deviation	N
0	14,4023	3,89303	517
1	18,9379	4,28162	483
Total	16,5930	4,67067	1000

Elegimos Descriptive statistics, Homogeneity tests, Spread vs. level plot, and Parameter estimates in the Display group → Continue OK.

Esta opción produce un análisis de covarianza para estimar el efecto del programa de participación, controlada por la variable salario antes de ingresar al programa.

La tabla 12.11 muestra una diferencia en la media del ingreso antes y después del programa; vemos una diferencia pequeña en las desviaciones estándar

El nivel Sig = 0.028 del test de Levene es menor a 0.05, lo que nos sugiere que el supuesto de varianzas iguales no se sostiene

Sin embargo, dado que hay sólo dos celdas definidas por la combinación de los niveles de factor, no se puede tener una conclusión definitiva al respecto

Tabla 12.11: Levene's Test of Equality of Error Variances[a]

Dependent Variable: Income after the program

F	df_1	df_2	Sig.
4,873	1	998	,028

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + incbef + prog

El Diagrama de puntos 12.4, de Spread (Desviación estándar) versus-level (Media) muestra lo que parece ser una relación entre la media y la desviación estándar. Pero, debido al escaso número de grupos, no se puede establecer nada concluyente.

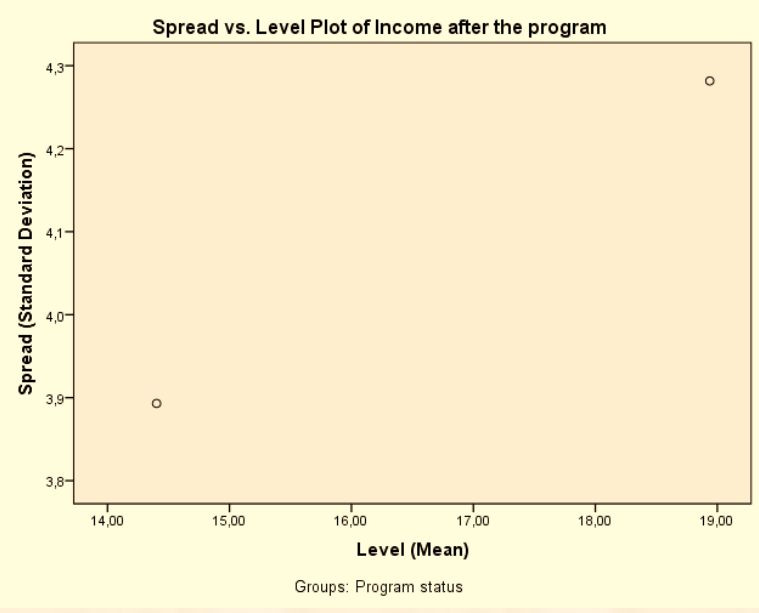


Diagrama 12.4

La diferencia en Spread (Desviación Estándar) es pequeña (0.38859) con relación a la diferencia en el nivel (Media) que es 4.5256

Así, podemos asumir que la varianza es la misma para ambos grupos.

Tabla 12.12: Tests of Between-Subjects Effects

Dependent Variable: Income after the program

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	12290,741[a]	2	6145,370	644,763	,000	,564
Intercept	131,400	1	131,400	13,786	,000	,014
incbef	7153,844	1	7153,844	750,571	,000	,429
prog	4735,662	1	4735,662	496,859	,000	,333
Error	9502,610	997	9,531			
Total	297121,000	1000				
Corrected Total	**21793,351**	**999**				

a. R Squared = ,564 (Adjusted R Squared = ,563)

En la tabla 12.12 el valor Sig = 0, 000 para las diferencias salariales entre incbef y prog

Ese valor es menor a 0.05, lo que nos indica que el programa tiene un influencia significativa sobre el ingreso

Los estimados de los parámetros se registran en la tabla 12.13 y muestran el efecto de cada predictor sobre el ingreso después del programa

Tabla 12.13: Parameter Estimates
Dependent variable: income after the program

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	**4,197**	**,556**	**7,548**	**,000**	**3,106**	**5,288**
incbef	**1,636**	**,060**	**27,397**	**,000**	**1,519**	**1,753**
[prog=0]	**-4,357**	**,195**	**-22,290**	**,000**	**-4,741**	**-3,974**
[prog=1]	**0[a]**	**.**	**.**	**.**	**.**	**.**

El valor -4.357 para (PROG = 0) es un dato muy importante, pues nos permite comparar el ingreso de dos sujetos antes del programa.

La comparación nos dice que después del programa, el que no participó del mismo tendrá un ingreso anual de $4357 dólares menos que para el que sí, participó

Resumen

Especificando una interacción entre la covariable y el factor, se puede testar la homogeneidad del parámetro de la covariable en todo el rango del factor.

Dado que el término de interacción no fue significativo en el ejemplo desarrollado, deducimos que los estimados de los parámetros de la covariable son homogéneos.

Por eso se procedió al análisis de covarianza

Así encontramos que la participación en el programa permitió el incremento del salario en un promedio de $4,357 con relación al que no participó

Si el término de interacción fuera significativo se podría usar el modelo con el término de interacción, en el entendido que al sopesar el efecto de la participación en el programa se complica con la presencia de la interacción.

Esto significa que cuando el término de interacción es significante, la diferencia entre las medias salariales de los participantes y no participantes cambia para valores diferentes de los niveles de ingreso antes del programa

El uso del procedimiento GLM Univariado para efectos aleatorios

En los análisis efectuados sobre la cadena de tiendas se examinó la relación entre los hábitos de compra de los clientes y el monto de gasto realizado en cada caso.

Hay, sin embargo, una gran variación entre tienda y tienda, que reduce la posibilidad de estimar los efectos de estos comportamientos.

Añadiendo la ubicación de las sucursales como un efecto aleatorio, se puede reducir la variación no explicada

Incrementamos así la exactitud de los estimados de los tér-minos de otros modelos.
La informacion para ese ejercicio está en el archivo *grocery_1month.sav.*

Usaremos el procedimiento GLM Univariado para ajustar un modelo con efectos fijos y aleatorios sobre los montos gastados en las compras.

> *Analyze → General Linear Model → Univariate → Reset* (restaura los valores)
> Elegimos *Amount spent* (Como variable dependiente)
>
> Elegimos *Who shopping for and Use coupons* (Como factores fijos)
> *Options → Estimates of effect size → Continue → OK.*

La tabla 12.14 tiene la información

Tabla 12.14

Tests of Between-Subjects Effects						
Dependent Variable: Amount spent						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1867084,001[a]	11	169734,909	37,796	,000	,551
Intercept	54651422,01	1	54651422,013	12169,66	,000	,973
shopfor	1329509,066	2	664754,533	148,026	,000	,466
usecoup	317508,903	3	105836,301	23,567	,000	,173
shopfor * usecoup	192031,603	6	32005,26	7,127	,000	,112
Error	1522377,820	339	4490,79			
Total	59475118,44	351				
Corrected Total	3389461,820	350				
a. R Squared = ,551 (Adjusted R Squared = ,536)						

El test de effectos between-subjets muestra que todos los términos del modelo tienen valores Sig = 0.000, menores que 0.05, son estadísticamente significativos.

Ahora añadiremos la sucursal ID como un factor de efectos aleatorios para ver si el modelo mejora o no

Caja de Diálogo GLM → Store ID (Como factor aleatorio) → Model.

Custom (Como modelo tipo) → *shopfor and usecoup in the Factors and Covariates list.*

Main effects from the Build Term(s) drop-down list and select the main effects to the model → trasladamos hopfor and use coup in the Factors and Covariates list.

Interaction from the Build Term(s) drop-down list and select the interaction term to the model para las mismas variables

Elegimos *storeid in the Factors and Covariates list → Build Term(s) drop-down list → Interaction → Continue → OK.*

El añadido de Store ID como un factor de efectos aleatorios reduce la variabilidad total no explicada, etiquetada como Error.

La reducción es de 1,522,377.82 (Tabla 12.14) a 1,073,908.57 (Tabla 12.15)
También reduce la variación explicada por los efectos principales: use coup y shpfor

Esto muestra que algunas de las variaciones explicadas originalmente por los términos de estos modelos pueden explicarse mejor con Store ID.

Tabla 12.15: Dependent Variable: Amount to spendSource		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	51402962,48	1	51402962,479	7173,14	,000
	Error	479937,37	66,97	7166,027[a]		
shopfor	Hypothesis	1109546,71	2	554773,357	144,646	,000
	Error	1073908,58	280	3835,388[b]		
usecoup	Hypothesis	253850,61	3	84616,869	22,062	,000
	Error	1073908,58	280	3835,388[b]		
shopfor * usecoup	Hypothesis	138871,54	6	23145,257	6,035	,000
	Error	1073908,58	280	3835,388[b]		
storeid	Hypothesis	448469,24	59	7601,174	1,982	,000
	Error	1073908,58	280	3835,388[b]		

Resumen

En este ejemplo, encontramos que añadiendo el factor aleatorio al modelo, incrementó la varianza relativa explicada por los otros términos del modelo

Procedimientos relacionados

El procedimiento GLM Univariado es útil para modelar relaciones lineales entre una variable dependiente de escala y una o más variables categóricas,pronosticadoras.

Si hay un solo factor, se puede usar One-Way ANOVA

Si hay covariadas es preciso usar el procedimiento Regresión Lineal, que es motivo del segundo volumen de esta obra.

13
LA CORELACIÓN LINEAL BIVARIADA

Las Correlaciones Bivariadas detectan las asociaciones entre pares que hay en un conjunto de variables, tal como veremos en la solución del siguiente problema

Es un procedimiento muy útil para determinar la fuerza y la dirección de la asociación entre variables de escala o variables ordinales.

Las asociaciones entre variables pueden dividirse en dos grandes grupos.

Primero, cuando ninguna de las variables es considerada como dependiente de otra.
Segundo, cuando una variable es claramente dependiente de otra u otras.

Si un alumno sobresale en la materia de matemáticas, es muy posible que también tenga buenas notas en estadística o viceversa.

No hay una clara relación de dependencia en esta asociación de variables.

Por el otro lado, el Consumo depende nítidamente del Ingreso; en este caso, el consumo es una variable dependiente del Ingreso, la que consideramos independiente.

En este capítulo analizaremos el primer caso bajo el nombre de Correlación Lineal Bivariada, es decir la asociación de dos variables.

La Correlación Lineal Bivariada

Es la asociación que existe entre dos variables cuando ambas varían, en sentido directo o inverso, pero no es posible establecer la dependencia de ninguna de ellas.

El grado de Correlación Bivariada es calculado por medio del Coeficiente de Correlación de Pearson, el estadístico rho de Spearman y el tau-b de Kendall.

Cada estadístico tiene sus respectivos niveles de significancia.

Antes de proceder a la estimación de los coeficientes de correlación, es muy útil tomar los valores respectivos de ambas variables y graficarlos.

Por otra parte, determinar la existencia de valores extremos (outliers) y evidenciar la existencia de una relación lineal.

Dos variables pueden estar muy correlacionadas entre sí, pero si la asociación no es lineal, el Coeficiente de Pearson no será útil para medir el grado de asociación.

Estudio de Caso

Con el objeto de incrementar las ventas, los expertos en diseño de vehículos de una firma han orientado su atención a los aspectos que el cliente considera importantes.

Por ejemplo, la importancia entre el gasto de combustible con relación a las ventas. Una manera de medir esta asociación es calcular la estimación entre ambas.

La información relativa al caso está en el archivo *car_sales.sav.*

Vamos a usas el procedimiento de la Correlación Bivariada para medir la importancia del consumo de combustible y las ventas de los vehículos.

Correlación Bivariada con el SPSS

Traemos el archivo *car_sales.sav*

Como un paso preliminar, averigüemos si las relaciones que podrían existir entre las diferentes variables que conforman el archivo.

Menú → Analyze > Correlate > Bivariate → *Sales in thousands y* Fuel efficiency como las variables de análisis → OK.

Tabla 13.1: Correlations

		Sales in thousands	Fuel efficiency
Sales in thousands	Pearson Correlation	1	-,017
	Sig. (2-tailed)		,837
	N	157	154
Fuel efficiency	Pearson Correlation	-,017	1
	Sig. (2-tailed)	,837	
	N	154	154

Traemos la tabla 13.1 de la Pantalla de Resultados; la relación entre la variables Sales in thousands y Fuel efficiency está medida por el Coeficiente de Pearson

El Coeficiente de Correlación de Pearson mide la asociación lineal entre dos variables de escala

En la tabla 13.1ese coeficiente es negativo -0.017(Asombroso) pero su valor Sig = 0.837 nos muestra que la relación no es significativamente diferente de cero.

El resultado sugiere que los expertos no deberían orientar sus esfuerzos en la fabricación de motorizados ahorradores de combustible.

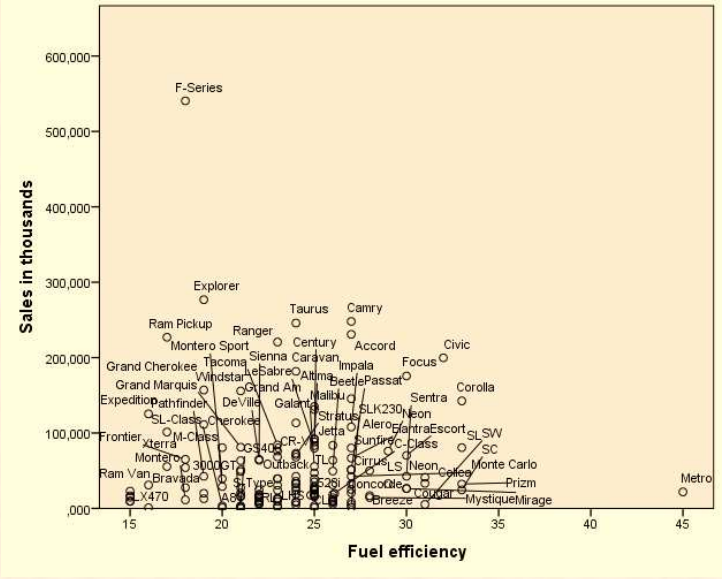
Es que esa variable no tiene un efecto apreciable en las ventas.

Pero el Coeficiente de Correlación de Pearson es más eficiente cuando las variables tienen una distribución por lo menos aproximada a la normal y no tienen "ouliers"

Un diagrama de puntos nos revelará los eventuales problemas
Para diseñarlo, vamos al Menú.

*Graphs → Chart Builder → Scatter/Dot gallery → Simple Scatter → Sales in thousands (en el eje de las Y's) → Fuel efficiency (*en las X's) *→ Groups/Point ID tab → Point ID Label*

Diagrama 13.1



Inmediatamente aparece, en la parte superior izquierda de la pantalla una casilla con la pregunta: ¿Point Label Variable?

De la lista de variables a la izquierda de la pantalla, elegimos *Model* y con el mouse la arrastramos hasta la casilla de la pregunta y tenemos el diagrama13.1

Esta última operación nos servirá para que cada uno de los puntos en el diagrama tenga su etiqueta mostrando el modelo de automotor al que representa.

Cada punto del diagrama tiene el nombre del modelo que representa
Pero notamos la presencia de dos outliers en el diagrama 13.1.

El primero, representado por el modelo F-Series, en la parte superior izquierda y el otro, representado Metro, en la parte inferior derecha del diagrama.

Dijimos ya que es necesario eliminar los valores extremos para lograr una mejor estimación del grado de correlación lineal entre dos variables.

Sin embargo, supondremos que la F-serie es importante para el equipo que diseña los modelos, por lo tanto lo mantendremos en la muestra.

Pero, el modelo "Metro", en la parte inferior derecha del diagrama es un valor extremo que, según los diseñadores, debemos excluir de la muestra.

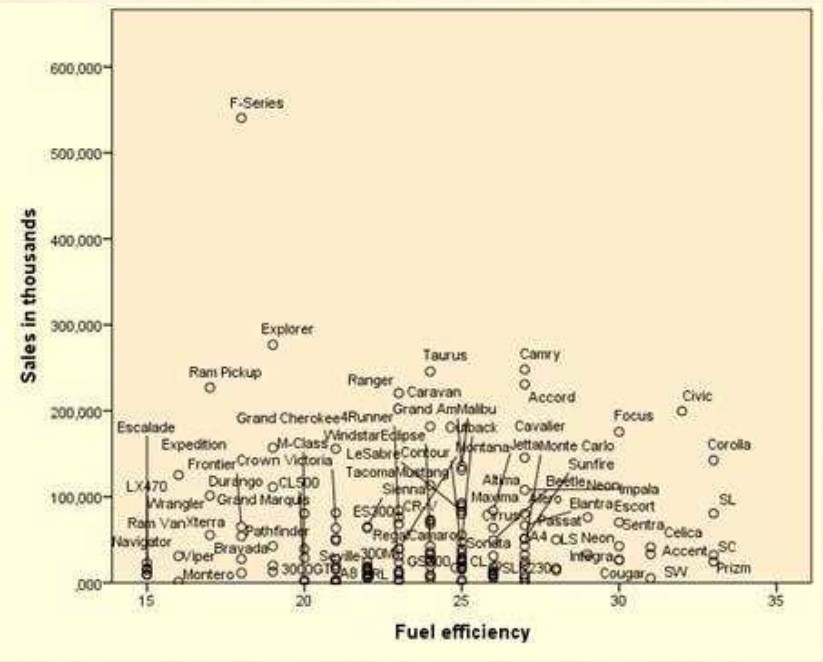
Menú → Data → Select Cases → If

Aparece una nueva pantalla; en la casilla de texto introducimos la variable *model* y le añadimos manualmente la expresión ~= 'Metro'.

Ese modelo ya no figurará en la muestra.

Para comprobarlo, repetimos el ejercicio y en la Pantalla de Resultados aparece el nuevo diagrama, al que denominamos Diagrama 13.2 sin el modelo "metro"

Diagrama 13.2



Análisis de los datos

Con la muestra filtrada por la omisión del modelo "metro" iniciamos el análisis de los datos, para ello vamos al Menú → Correlate → Bivariate

De la lista de variables a la izquierda de la pantalla traemos Fuel efficiency → Log-transformed sales como variables de análisis.

Tabla 13.2 Correlations

		Fuel efficiency	Log-transformed sales
Fuel efficiency	Pearson Correlation	1	,136
	Sig. (2-tailed)		,093
	N	153	153
Log-transformed sales	Pearson Correlation	,136	1
	Sig. (2-tailed)	,093	
	N	153	156

Removido outlier "metro" y analizando la variable log-transformed sales vemos que la correlación es positiva 0.136 pero aún no es significativamente diferente de 0.

Sin embargo es posible deducir que el mercado para camiones y automóviles son diferentes y las razones para comprar uno u otro no son las mismas

Para salir de dudas diseñaremos otro diagrama de puntos, pero de manera tal que haya una variable diferenciadora, en este caso, el tipo de vehículo.

Para producir el diagrama de puntos para las variables Log-transformed sales y Fuel efficiency, controladas por la variable vehicule type hacemos lo siguiente.

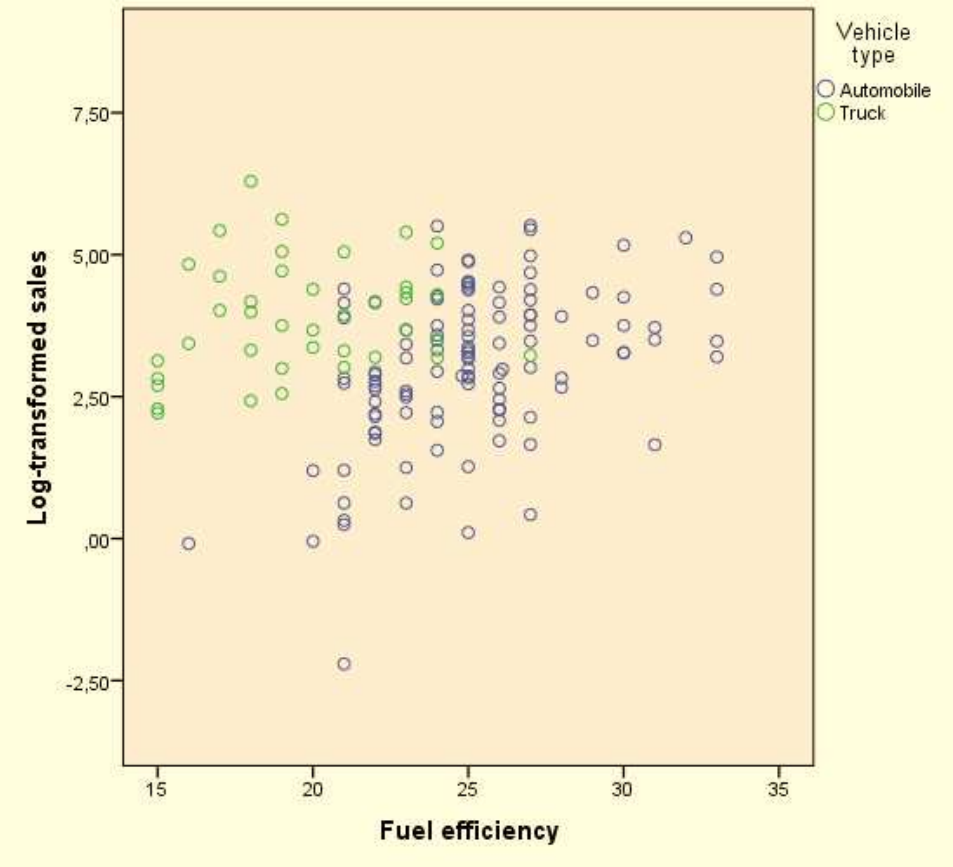
Chart Builder → Groupe Scatter → Log-transformed como la variables de las Y's
Fuel efficiency para el eje de las X's

Vehicule type como la variable que define los colores →OK.

El diagrama de puntos 13.3 muestra camiones y automóviles con diferentes colores. Esta división puede mejorar el nivel de Asociación entre las variables.

Después de la transformación logarítmica de la variable Sales in thousands el outlier de la parte superior izquierda del diagrama también ha desaparecido

Diagrama 13.3



Ahora debemos dividir los datos originales de acuerdo con el tipo de vehículo

Menú → Data → Split File → Compare groups → Select Vehicle type como variable que servirá referente para la division de grupos → OK.

Para analizar el archivo con los datos divididos en grupos
Menú → Correlate → Bivariate → OK.

La Pantalla de Resultados registra la tabla 13.4; allí vemos que la división de los datos del archivo nos permite una asociación más clara.

En efecto, el Coefficiente de Pearson para la correlación entre Fuel efficiency y Log transformed sales controlada por automóviles es 0.451 y su valor Sig es 0.000.

Para los camiones, el Coeficiente de Pearson es0.203, pero su valor Sig es 0.210 lo que nos sugiere que no hay una correlación entre camiones y fuel effciency

De todos modos, este ejercicio nos sirvió para establecer algunas maneras que nos permitan mejorar nuestros hallazgos transformando variables usando medidas de correlación no paramétricas

Tabla 13.4

Correlations

Vehicle type			Fuel efficiency	Log-transformed sales
Automobile	Fuel efficiency	Pearson Correlation	1	,451**
		Sig. (2-tailed)		,000
		N	113	113
	Log-transformed sales	Pearson Correlation	,451**	1
		Sig. (2-tailed)	,000	
		N	113	115
Truck	Fuel efficiency	Pearson Correlation	1	,203
		Sig. (2-tailed)		,210
		N	40	40
	Log-transformed sales	Pearson Correlation	,203	1
		Sig. (2-tailed)	,210	
		N	40	41

**. Correlation is significant at the 0.01 level (2-tailed).

Otros estadísticos de Correlación

Los estadísticos Spearman's rho y Kendallos tau-b miden el orden de rango de las asociaciones entre dos variables de escala u ordinales.

Tabla 13.5

Correlations

	Vehicle type			Fuel efficiency	Log-transformed sales	Sales in thousands
Spearman's rho	Automobile	Fuel efficiency	Correlation Coefficient	1,000	,425**	,425**
			Sig. (2-tailed)	.	,000	,000
			N	113	113	113
		Log-transformed sales	Correlation Coefficient	,425**	1,000	1,000**
			Sig. (2-tailed)	,000	.	.
			N	113	115	115
		Sales in thousands	Correlation Coefficient	,425**	1,000**	1,000
			Sig. (2-tailed)	,000	.	.
			N	113	115	115
	Truck	Fuel efficiency	Correlation Coefficient	1,000	,237	,237
			Sig. (2-tailed)	.	,141	,141
			N	40	40	40
		Log-transformed sales	Correlation Coefficient	,237	1,000	1,000**
			Sig. (2-tailed)	,141	.	.
			N	40	41	41
		Sales in thousands	Correlation Coefficient	,237	1,000**	1,000
			Sig. (2-tailed)	,141	.	.
			N	40	41	41

**. Correlation is significant at the 0.01 level (2-tailed).

Su ventaja estriba en que no toman en cuenta la distribución de las variables, mientras que el Coeficiente de Pearson exigía una distribución cercana a la Normal.

Para lograr un análisis usando Spearman's rho, vamos a la caja de diálogo Correlación Bivariada → Sales in thousands como una variable de análisis

Desactivamos el botón de Pearson y activamos el de Spearman → OK.

La pantalla de resultados nos trae la tabla 13.5 en la que los informes de Spearman's rho están divididos en automóviles y camiones.

En la tabla comprobamos que la asociación entre Log-transformed sales and Fuel efficiency es significativa

Spearman's rho tiene la misma correlación con las ventas no transformadas

Esto se debe a que el estadísticos se basa en órdenes de rango, los que no son cambiados por la transformación logarítmica

De esta manera, los valores extremos, outliers, tienen un efecto menor que en el estadístico Spearman's rho, por lo que es útil como medida de asociación.

Resumen

El uso de la Correlación Bivariada nos proporcionó una correlación negativa entre sales in thousands y fuel efficiency, lo que no tenía sentido.

Luego de remover los outliers y transformar la variable sales in thousands en sus logaritmos, la correlación se hizo positiva, aunque no significativamente diferente de 0

Separando los datos en camiones y automóviles encontramos una correlación positiva estadísticamente significativa entre ventas y fuel efficiency para automóviles.

También llegamos a los mismos resultados sin necesidad de transformar las variables, usando Spearman's rho.

Esto nos permite preguntarnos ¿Por qué debemos transformar las variables si el Estadístico Spearman's rho es tan conveniente.

Si bien los órdenes de rango son efectivos para detectar alguna clase se asociación entre 2 variables, necesitamos una transformación para que la relación sea lineal.

Esto se debe a que hay más modelos pronosticadores disponibles para relaciones lineales, además de que son más fáciles para implementar e interpretarlos.

Correlaciones Parciales

Las Correlaciones Parciales calculan los coeficientes parciales de correlación que describen relaciones lineales entre dos variables

Pero, lo hacen mientras se controla el efecto de una o más variables adicionales; todas las variables deben ser numéricas (escala)

Caso de estudio

Los estudios estadísticos de un programa de ayuda a la salud con fondos gubernamental revelaron algo insólito: crecen los fondos y la tasa de mortalidad también.

Al parecer, según los datos, que son correctos, la población estaría mejor sin los fondos adicionales que el gobierno otorga a hospitales y clínicas.

Esto significaba que había una correlación positiva entre el incremento de fondos para la salud y el aumento de las tasas de mortalidad.

Aquí es necesario poner en evidencia que antes de correr un programa estadístico se debe contar con una teoría que respalde lo que se busca.

En este caso, los números contradicen una teoría razonable, pues no es posible pensar que un incremento de fondos para la salud aumente la tasa de mortalidad.

Para averiguar que sucede vamos a realizar un análisis de Correlación sobre el archivo health_funding.sav

Con el archivo en la pantalla: *Analyze → Correlate → Partial → Health care funding → Reported disease rate* como las variables

Ahora incluimos una nueva variable: *Visits to health care providers* que la introducimos en la casilla inferior de la pantalla abierta como la variable de control

→ Options → Zero-order correlations → Continue → OK.

La Pantalla de resultados registra la tabla 13.6, que. muestra las correlaciones de orden cero, de las tres variables, sin control de variable alguna

También la correlación parcial de las dos primeras controladas por los efectos de la tercera variable

La correlación de orden cero entre *health care funding y disease rates* cuando no introducimos la variable control es 0.0737 para un valor Sig = 0.000

Esto que nos muestra que hay una correlación significativa entre ambas variables.

Pero, la correlación parcial de la variable que controla, *visits to health care*, es insignificante; el valor de la correlación es 0.013 y el valor Sig = 0.928

Esto significa que la correlación entre las variables *health care funding y disease rates* es casi nula cuando introducimos la variable de control *visits to health care*

Tabla 13.6

Correlations

Control Variables			Health care funding (amount per 100)	Reported diseases (rate per 10,000)	Visits to health care providers (rate per 10,000)
-none-[a]	Health care funding (amount per 100)	Correlation	1,000	,737	,964
		Significance (2-tailed)	.	,000	,000
		df	0	48	48
	Reported diseases (rate per 10,000)	Correlation	,737	1,000	,762
		Significance (2-tailed)	,000	.	,000
		df	48	0	48
	Visits to health care providers (rate per 10,000)	Correlation	,964	,762	1,000
		Significance (2-tailed)	,000	,000	.
		df	48	48	0
Visits to health care providers (rate per 10,000)	Health care funding (amount per 100)	Correlation	1,000	,013	
		Significance (2-tailed)	.	,928	
		df	0	47	
	Reported diseases (rate per 10,000)	Correlation	,013	1,000	
		Significance (2-tailed)	,928	.	
		df	47	0	

a. Cells contain zero-order (Pearson) correlations.

Las tasas de mortalidad parecen crecer con el incremento de los fondos de ayuda a la salud debido al crecimiento del número de personas con acceso a esos servicios

Por esta razón, los administradores de los hospitales y otros centros de atención médica informan sobre el número creciente de muertes

En consonancia con el mayor número de personas enfermas que ahora acuden a los centros de salud.

Lo importante de este ejercicio es que nos instruye sobre la necesidad de establecer una teoría previa antes de realizar cualquier programa estadístico.

Resumen

Las Correlaciones Parciales son apropiadas sólo para variables numéricas.

Si se tiene variables categóricas usamos el procedimiento Crosstabs, allí las variables "de capa" son similares a las variables de control que vimos en este apartado.

El Procedimiento Crosstabs

Crosstable (Tabulación Cruzada) es una técnica para examinar las relaciones entre dos variables categóricas, eventualmente controladas por variables "de capa"

Este procedimiento testa la independencia y mide la asociación para datos nominales y se puede obtener estimados sobre el riesgo relativo de un evento, dada la presencia o no de una característica particular y testar diferencias significativas

Caso de Estudio

Con el objeto de establecer los grados de satisfacción de la clientela, una firma realizó una encuesta de 582 clientes en 4 sucursales diferentes

La encuesta reveló que la calidad del servicio al cliente fue el más importante factor en el marco de un grado de satisfacción general.

Con esta información, se desea testar si cada sucursal provee un servicio similar y adecuado al cliente; los resultados de la encuesta están en el archivo *satisf.sav.*

Vamos a usar el procedimiento Crosstabs para testar la hipótesis de que los niveles de satisfacción del cliente son constantes en todas las sucursales de la firma.

Analyze → Descriptive Statistics →> Crosstabs → Store como la variable de fila→ *Service satisfaction* como la variable de columna → *Statistics.*

Elegimos *Chi-square → Contingency Coefficient → Phi and Cramer's V → Lambda → Uncertainty coefficient → Continue → OK.*

Tabla 13.7

Store * Service satisfaction Crosstabulation

		Service satisfaction					Total
		Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	
Store	Store 1	25	20	38	30	33	146
	Store 2	26	30	34	27	19	136
	Store 3	15	20	41	33	29	138
	Store 4	27	35	44	22	34	162
Total		93	105	157	112	115	582

La tabla 13.7 tiene la información sobre la frecuencia de respuesta en cada sucursal.

Si cada sucursal provee un nivel de servicio similar, el patrón de respuestas será similar en todas las sucursales.

En la *tabla 13.7* vemos que la mayoría de las respuestas ocurren entre Somewhat Negative Neutral y Somewhat Positive

Pero la sucursal 2 parece tener un número menor de clientes satisfechos.
En cambio la sucursal 3 parece tener un número menor de clientes insatisfechos.

La tabla 13.7, por sí sola, no nos permite saber si estas diferencias son reales o no dadas las variaciones aleatorias que podrían existir.

Para asegurarnos traemos de la Pantalla de Resultados la tabla 13.7, que tiene el test chi-square

El Estadístico chi-square testa las medidas de discrepancia entre los datos de las celdas y lo que se esperaría si las filas y las columnas no estuvieran relacionadas.

En la tabla 13.8 el valor Sig de 2 colas para el nivel de significancia asintótica (The two-sided asymptotic significance) del estadísticos chi-square es mayor que 0.05

Con ese dato, podríamos asumir que las diferencias son debidas variaciones aleatorias y que todas las tiendas tendrían el mismo nivel de atención al cliente.

Sin embargo, no todos los clientes que respondieron a la encuesta tuvieron contacto con los representantes de servicio al cliente.

Tabla 13.8 Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16,293[a]	12	,178
Likelihood Ratio	17,012	12	,149
Linear-by-Linear Association	,084	1	,772
N of Valid Cases	582		

a. 0 cells (0,0%) have expected count less than 5.

b. The minimum expected count is 21,73.

Ahora clasificaremos los clientes de todas las sucursales en dos grupos: los que tuvieron y los que no tuvieron contacto con el servicio de atención al cliente.

Para hacerlo, vamos otra vez a la caja de diálogo de Crosstabs.
Elegimos Contact with employee como la variable capa → OK

La tabla que obtenemos, 13.9, divide la anterior en dos partes

Ahora podemos ver que los clientes que no tuvieron contacto con el encargado de servicio al cliente son agrupados

Tabla 13.9 Store * Service satisfaction * Contact with employee Crosstabulation

Count

Contact with employee			Service satisfaction					Total
			Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	
No	Store	Store 1	16	9	18	17	19	79
		Store 2	2	15	16	13	12	58
		Store 3	9	14	23	22	14	82
		Store 4	17	14	19	10	10	70
	Total		44	52	76	62	55	289
Yes	Store	Store 1	9	11	20	13	14	67
		Store 2	24	15	18	14	7	78
		Store 3	6	6	18	11	15	56
		Store 4	10	21	25	12	24	92
	Total		49	53	81	50	60	293
Total	Store	Store 1	25	20	38	30	33	146
		Store 2	26	30	34	27	19	136
		Store 3	15	20	41	33	29	138
		Store 4	27	35	44	22	34	162
	Total		93	105	157	112	115	582

Al parecer habría una asociación significativa entre las tiendas 2 y niveles bajos de atención al cliente; para estar seguros, comprobamos con el test chi-square.

El test chi-square, Tabla 13.10, realiza separadamente para los clientes que tuvieron contacto con el encargado de los servicios al cliente y los que no tuvieron ese contacto.

Tabla 13.10

Chi-Square Tests

Contact with employee		Value	df	Asymp. Sig. (2-sided)
No	Pearson Chi-Square	20,898[b]	12	,052
	Likelihood Ratio	22,937	12	,028
	Linear-by-Linear Association	3,514	1	,061
	N of Valid Cases	289		
Yes	Pearson Chi-Square	25,726[c]	12	,012
	Likelihood Ratio	25,777	12	,012
	Linear-by-Linear Association	1,993	1	,158
	N of Valid Cases	293		
Total	Pearson Chi-Square	16,293[a]	12	,178
	Likelihood Ratio	17,012	12	,149
	Linear-by-Linear Association	,084	1	,772
	N of Valid Cases	582		

El valor Sig = 0.052 para los clientes que no tuvieron contacto con el servicio de atención al clientes; es sugestivo pero no concluyente.

Aunque parece haber alguna relación entre la sucursal y el servicio de satisfacción al cliente, el valor Sig está en la "frontera" de aceptación y rechazo de la hipótesis.

Para estar más seguros sería necesario un análisis separado para esos clientes y así determinar si hay algún otro factor que influye en la relación.

El nivel Sig del test para los clientes que tuvieron contacto con el empleado del servicio de atención al cliente es 0.012, menor al valor 0.05.

Sobre esa base podemos concluir que la relación observada en la tabla 13.3 no se debía al azar, más bien, era significativa.

El test chi-square es muy útil para determinar si hay o no una relación entre dos variables, no nos dice algo sobre la fuerza de esa relación.

Para ello vamos a recurrir a las medidas simétricas.
Los datos simétricos están registrados en la tabla 13.11

Miden separadamente los estadísticos de los grupos de clientes que tuvieron y que no tuvieron contacto con los empleados del servicio de atención al cliente.

Esas medidas se basan en el estadístico chi-square.

El estadístico Phi es el resultado de la razón del estadístico chi-square y el número total de observaciones, debidamente ponderadas.

Es la más "optimista" de las medidas simétricas y a diferencia de la mayor parte, no tiene un límite superior cuando las variables tienen más de dos categorías.

Tabla 13.11

Symmetric Measures

Contact with employee			Value	Approx. Sig.
No	Nominal by Nominal	Phi	,269	,052
		Cramer's V	,155	,052
		Contingency Coefficient	,260	,052
	N of Valid Cases		289	
Yes	Nominal by Nominal	Phi	,296	,012
		Cramer's V	,171	,012
		Contingency Coefficient	,284	,012
	N of Valid Cases		293	
Total	Nominal by Nominal	Phi	,167	,178
		Cramer's V	,097	,178
		Contingency Coefficient	,165	,178
	N of Valid Cases		582	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

El estadístico Cramer's V lleva al estadístico phi hasta su máximo valor posible, que es siempre 1; a medida que el número de columnas se incrementa.

Cramer's V se hace más conservador con relación a Phi.

El Coeficiente de Contingencia toma valores entre 0 y la raíz cuadrada de (k-1)/k expresión en la que k es igual al número de filas o de columnas.

Si el número de filas es menor que el de columnas, tomará el número de filas; si el número de columnas es menor que el de filas, tomará las columnas como referente

El Coeficiente de Contingencia se vuelve más conservador con respecto a phi a medida que la asociación entre las variables es más fuerte.

El valor Sig para las tres medidas en 0.012, lo que indica que hay una relación estadísticamente significativa entre las variables testadas

Pero, los valores de los tres estadísticos están por debajo de 0.3, por lo que si bien la relación no se debe al azar, no es muy fuerte

Mientras estas medidas nos dan ciertas pautas sobre la fortaleza de la asociación entre variables, no nos ofrecen una interpretación intuitiva

Hay otras medidas de interés que serán analizadas en el segundo tomo de esta obra. Con esto terminamos el Manual de ESTADÍSTICA APLICADA CON SPSS

Este libro fue distribuido por cortesía de:

Para obtener tu propio acceso a lecturas y libros electrónicos ilimitados GRATIS hoy mismo, visita:
http://espanol.Free-eBooks.net

Comparte este libro con todos y cada uno de tus amigos de forma automática, mediante la selección de cualquiera de las opciones de abajo:

Para mostrar tu agradecimiento al autor y ayudar a otros para tener agradables experiencias de lectura y encontrar información valiosa, estaremos muy agradecidos si
"publicas un comentario para este libro aquí".

INFORMACIÓN DE LOS DERECHOS DEL AUTOR

Free-eBooks.net respeta la propiedad intelectual de otros. Cuando los propietarios de los derechos de un libro envían su trabajo a Free-eBooks.net, nos están dando permiso para distribuir dicho material. A menos que se indique lo contrario en este libro, este permiso no se transmite a los demás. Por lo tanto, la redistribución de este libro sín el permiso del propietario de los derechos, puede constituir una infracción a las leyes de propiedad intelectual. Si usted cree que su trabajo se ha utilizado de una manera que constituya una violación a los derechos de autor, por favor, siga nuestras Recomendaciones y Procedimiento de Reclamos de Violación a Derechos de Autor como se ve en nuestras Condiciones de Servicio aquí:

http://espanol.free-ebooks.net/tos.html